Application of Dimension Reduction and Clustering Methods for Detection of Faulty Operations in Process Systems

By

MELISA MOLLAIAN
THESIS

Submitted in partial satisfaction of the requirements for the degree of

MASTER OF SCIENCE

in

Chemical Engineering

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Ahmet Palazoglu, Chair

Nael H. El-Farra

Matthew J. Ellis

Committee in Charge

2021

Acknowledgements

Advancing through graduate school is a unique experience and was very different from what I anticipated. My experience was also especially affected by COVID-19 pandemic and I believe proceeding in these two years would not have been possible without the help of many people who supported me.

First and foremost, I would like to thank my advisor, Professor Ahmet Palazoglu for his guidance, mentorship and invaluable advice. Only with his continuous support I was able to explore this research direction and develop my skills. I deeply appreciate his patience and motivation and the opportunities provided by working in his research lab. I would also like to thank Professor El-Farra and Professor Ellis for serving on my thesis committee.

I am also grateful for the help I received from my colleagues and many others in graduate school, most importantly Dr. Gyula Dörgő. I am grateful for his friendship and support while working on projects and for helping me grow as a researcher and an engineer.

Last but not least, I would like to express my sincere gratitude to my parents and my friends without whose support I never could have reached where I am. I am thankful for your love and for tolerating me over my tough moments in the past two years.

I am grateful for all I learnt in life, and therefore I dedicate this thesis to all those who are deprived of educational rights; Wishing the future would hold equal opportunities for humans all over the world.

Abstract

Widespread application of distributed control systems and measurement technologies in chemical plants are prerequisites for quality control and safety monitoring of processes. This consequently has prompted large-scale data acquisition and storage from different processes and various sectors of a plant. The collected data holds information about the behavior of the process which can further assist in developing data-driven methods for detection and diagnosis of process anomalies (faults). Considering the recognition data-driven modeling has received in the past decades, additional studies on incorporation of data mining techniques for knowledge discovery from chemical process data would seem to be a useful practice.

Data mining methods can be used to identify different groups or classes present in a dataset, or in the context of process systems engineering, distinguish faulty states from normal process operation in historical datasets. Two key tools used in this practice are dimension reduction techniques and clustering methods. The former helps with feature extraction from the data and the latter detects groups within the data, therefore, a productive combination of these two tools can be promising in facilitating the fault detection and diagnosis applications. The first part of this research studies the performance of different combinations of dimension reduction techniques and clustering methods to evaluate their ability in detection process faults, and demonstrates the higher compatibility of some of these methods with others.

While performing clustering, detecting the number of clusters present in the dataset is either a direct aim of the study, or it can be beneficial in choosing the most suitable parameters and/or labelling. Motivated by the first part of the research, performance of clustering methods on a dataset before and after different dimensionality reduction techniques is studied using internal metrics for clustering performance. Based on multi-objective optimization, an approach is proposed to detect the cluster numbers in an unsupervised manner which successfully presents the expected cluster numbers in three distinct applications.

In summary, this research aims to assist data-driven fault detection practices in chemical processes by elucidating the synergy between two data mining techniques; dimension reduction and clustering methods. Also, the introduced approach to detect the expected cluster number in a dataset contributes to the progress of unsupervised state-isolation studies for any application and generic datasets.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

Computers, internet, and more importantly, technology, have played an undeniable role in the evolution of data storage and data processing, impacting personal, social and corporate environments. Now, in almost every aspect of life, data is being extracted and stored with the goal to improve the current structures and strategies. With the rise of Internet of Things (IoT), systems are now able to access raw data from various sources over a network and analyze this information to extract knowledge [1]. Big Data has found its place in diverse applications in such a manner that by utilizing the potential of data availability, paths to process/product improvement can be discovered. Accomplishments such as recommendation engines, speech recognition and image classification are just some examples of significant applications that became possible due to abundance of data and advancement of technology [2]. To take advantage of this potential provided by data storage, one important matter after gathering the relevant data is the availability of tools to study them [3].

The process of studying gathered data is closely related to the data characteristics. Data properties such as volume, variety and velocity are defining features in determining the manner in which the data can be approached and analyzed [1]. Since data gathered from different domains are inherently different, generalizing courses of action for such analysis can be a difficult task. Evidently, many opportunities arise from the challenges posed by the development of information extraction. Furthermore, data mining techniques emerged to extract information, patterns, behaviors and in general, knowledge from the datasets and has evolved as database and information technology proceeded systematically. Data mining can be viewed as one of the steps of Knowledge Discovery from Data (KDD), which includes other steps such as data selection and knowledge presentation [3], or can be viewed as the KDD itself. Various techniques from other domains have also contributed to data mining in order to develop models to handle different data characteristics; some examples of these domains are statistics and visualization. Data mining has a deep-rooted connection with statistics due to its data-driven approach, and many methods such as regression models and Bayesian networks are now widely used for data mining tasks. Integration of visualization in data mining process is also explored frequently, since in application of visualization, data are treated as objects and can be represented in projections comprehensible to humans for examination steps [4].

A dataset might contain numerous hidden patterns, therefore different types of functions, techniques and algorithms can be used to extract information. Data mining functions are concerned with the type of patterns which can be extracted during the process, and some major data mining functions are classification, clustering and summarization, for which by using different data mining techniques such as machine learning and neural networks, a selection of algorithms would be available to work with in the data mining process [4]. While as a data mining technique, machine learning applications might use the same algorithms as other data mining applications, the main focus of machine learning is to adapt the algorithms in such a way to make predictions for future datasets, however, data mining is used to extract knowledge from existing data. Moreover, data mining techniques can be classified into three categories of supervised learning, unsupervised learning and reinforcement learning. Supervised learning utilizes a labeled dataset for training purposes to learn the patterns for classification and prediction purposes, whereas in unsupervised learning, data is not labeled and the method aims to discover and learn the classes and labels, such as in clustering. Reinforcement learning uses data and some critic for a given objective, to learn the relationship with critic while self-optimizing [3]. Hence, due to the capabilities of data mining, it has become an essential task in various domains of research, such as medical, environmental and traffic control [5, 6, 7]. Various methods and algorithms have been introduced throughout the years, and are continuing to improve and develop. However, the data mining process is not straightforward. Processing and preparation of the data is a challenging task and selection of an algorithm highly depends on the application and data characteristics.

One of the applications of data mining is in chemical engineering process systems, as it is an excellent domain by having two important features required for data mining applications: data availability and opportunities for enhancement. Chemical process systems are immensely complex and require constant monitoring of all variables throughout a chemical plant to ensure the safety of the plant and prevent accidents which may be caused by faults and abnormal operational states. Industrial statistics showed that about 70% of industrial accidents are caused by human errors [8]. Although recent developments are improving these statistics, there are still accidents that result in injuries and incur costs [9]. Therefore, ensuring a reliable control system to observe the behavior of different plant sections, as well as the overall process is vital.

A fault (or an abnormal operational state) can be defined as an unacceptable deviation from at least one characteristic property of the system, which can happen regularly if a well-founded control system is not in place. Constant monitoring and measuring of process variables in the system via sensors will help in detecting the faults and anomalies in a timely manner to improve productivity and increase safety. This is usually done by comparing the process state with the normal process behavior, and setting alarms based on the expected values for different measurements [10]. Moreover, process information obtained from the fault is crucial in returning the system to normal operation and also, it can be used to further analyze the nature of the fault [11]. In the end, by storing all the acquired data, the process monitoring strategies result in massive databases containing the process behavior. These large datasets constitute the basis of fault detection and diagnostics framework and future control strategies for chemical plants [12].

Fault detection methods can be classified into three main categories: model-based methods [13], knowledge-based methods [14], and process history-based methods [8]. The first category of methods, models are developed to observe deviations and require the process system models to be as exact as possible in order to detect faults. These models can be difficult to obtain and indeed, not entirely possible in real-life cases, due to system complexity and development cost. The second category, knowledge-based methods, use the captured knowledge from observations from the process behavior to develop a qualitative model for fault detection and diagnostics. In contrast to these two approaches where information about the process is required to create a model, in the third category, large amount of historical data from the process is sufficient for developing fault detection methods. One of the benefits of the third category compared to modelling approaches other than ease of implementation, is their flexibility in utilizing a variety of data types [9].

Chemical plants use distributed sensor networks for monitoring the operation of the system, and therefore large amounts of process data can be recorded and used for developing data-driven methods [10]. Therefore, data mining and machine learning approaches can offer solutions for fault detection challenges by utilizing the available data. Thus, data-driven approaches have caught much attention in research and industry. A comprehensive review on the historical-based methods for fault detection and diagnosis is available in the work of Venkatasubramanian et al. [8] where they discuss relative strengths and weaknesses of different approaches. Data-driven process monitoring is also a subject of research, and it is called statistical process monitoring (SPM). It benefits from the use of multivariate statistics and machine learning methods as well. Due to the nature of these methods, they are easier to use for fault detection and fault identification, and require less a priori knowledge compared to other monitoring approaches [10, 15].

Data-driven methods essentially contemplate the fault detection process as a pattern recognition problem in order to classify the data points and identify different states and therefore, recognize possible faults. The ultimate goal is to extract major trends, which can also be beneficial in process design [8]. Data-driven fault detection methods can be categorized into two main groups, supervised and unsupervised approaches. Many important fault detection methods in chemical engineering use supervised learning strategies such as neural networks [16] and support vector machines [17] meaning groups of labeled data should be available for at least training purposes. While in some cases labeled data from plants might be available, properly labeling data from long-term operations is a time consuming task, assuming the operational state of different sensors is known. Moreover, this labeling requires much familiarity with measurements and different operating regimes of a plant [18]. However, using unsupervised learning strategies requires less a priori information about the data, if any. Therefore, these methods can be more favorable in separating the faulty data from normal data. One important example of the unsupervised learning approaches is principal component analysis [19]. Each category of data-driven fault detection approaches have their advantages and disadvantages and no single method is applicable in all cases. Nevertheless, due to the increasing complexity of process systems and their performance and the constant need for improvement, there is always a demand for development of new methods.

1.2 Research motivation

Unsupervised learning strategies help discovering hidden patterns and behaviors within the data, and two approaches of these strategies are clustering methods and dimension reduction techniques. In simple terms, for clustering methods, the aim is to identify the number of clusters in a dataset and extract classes or groups. For dimension reduction techniques, the purpose is to extract information from high-dimensional data and represent them in lower number of dimensions [18]. These methods fit into the category of unsupervised learning, however, they can also be used for training purposes if the data is labeled. Nevertheless, in utilization of any data mining method, an extent of knowledge is required about the dataset, the application and available methods to select the best one, and this process can be time-consuming for non-(domain) experts in data science.

Many studies have been performed to help select the best matching method for different purposes. For example, Espadoto et al. [20] compared 44 dimension reduction techniques for 18 datasets and 7 quality metrics to answer the question how to choose the best technique for a given context. In a work done by Caruana et al. [21], after discussing the difficulty of evaluating the clustering performance, they present an approach in which instead of finding one optimal clustering, the user examines a small number of clusterings to decide which clustering is the most useful one. However, these comparison studies only consider individual applications of different data mining techniques, whereas in many cases a combination of these techniques can be more useful for information extraction and machine learning purposes. Each data mining method can be useful for discovering specific patterns, and while studying the performance of their combinations can be difficult, it can lead to new guidelines on how to extract information more efficiently and more accurately. There are studies such as the work done by Tang et al. [22] who have compared the performance of a dimension reduction technique when combined with clustering methods for text clustering purposes. Another work performed by Thomas et al. [18], studies combination of 5 dimension reduction techniques and 4 clustering methods for industrial chemical process data. Altogether, performing a comprehensive study which covers comparison of all combinations of categories of clustering methods and dimension reduction techniques for chemical engineering purposes seems promising, and is the motivation for the first part of this research.

Many data mining methods require hyperparameter selection or initialization. After the appropriate method has been selected, the next step is to find the suitable parameters. The results of the methods highly depend on these parameters, and their determination is a difficult task. Parameters may be data-specific and the selection requires a deep understanding of the data [3]. For some clustering methods, such as the well-known k-means, the required parameter to perform the clustering is the number of clusters. This information is rarely available; indeed, one of the main objectives of clustering usually is to find the number of classes present in the data, as this number is an important characteristic of the dataset. Hence, one important research direction is to facilitate parameter tuning of the efficient algorithms. Studies have been performed to develop solutions that require less effort from the user in the process. For example, Belkina et al. [23] introduced automated optimized parameters for t-distributed Stochastic Neighbor Embedding (opt-SNE) for automatically finding t-SNE parameters via fine-tuning. Their approach eliminates the need for trial and error runs for empirically finding the t-SNE parameters and allows faster t-SNE implementations. As another example, Beaver et al. [24] presented a graphical solution to determine the appropriate range for the number of clusters, after introducing aggregated k-means to overcome the shortcomings of traditional clustering algorithms.

In cases of high-dimensional data, detection of the number of clusters becomes more challenging due to the high number of measurements and increased level of correlations among them. Therefore, finding the appropriate number of clusters, apart from the clustering analysis, needs to be specifically addressed. Various studies have provided approaches to determine the number of clusters, and a comprehensive review of them is presented in a survey done by Hancer and Karaboga [25]. Based on their comparison, approaches to find the cluster number fit into one of three categories: traditional approaches, merge-split based approaches and computation based approaches. In one type of traditional approach, it is assumed that by changing the parameter of cluster number and fixing all other parameters, there is a significant change in an evaluation graph. The graph shows the changes in a validity index versus the number of clusters, and the significant change in the evaluation parameter is concerned with the correct number of clusters, i.e., the knee point [26]. In another type of traditional approach, the detection is performed using resampling of the data, assuming different copies of data should provide similar results in terms of number of clusters [27]. Merge-split based approaches try to split or merge clusters iteratively according to some criteria. These criteria are either based on statistical rules, as seen in X-means [28] or are homogeneity based, i.e., based on cluster-related distances, where in the latter, split-merge thresholds need to be specified by the user. An example of homogeneity based methods is ISODATA [29]. Computation based methods are widely applied to evolutionary computation (EC) problems with different encoding schemes, and are placed in two groups of single-objective and multi-objective approaches. In this category, one or multiple validation metrics are optimized in order to find the optimum number of clusters. Although single-objective approaches contributed to the advancement of detecting the cluster number and many algorithms were introduced such as GCUK [30] and CGA [31], a single objective may not be enough to correctly judge the structure of the data. Thus, multi-objective optimization approaches try to simultaneously optimize more than one validity criterion and incorporate different characteristics of the data. There are several studies carried out by computation-based problems utilizing multi-objective optimization to detect cluster numbers [25], and they have performed well in terms of clustering quality and number of clusters due to their global approach. This research direction has potential and studies on applications of multi-objective clustering in fields other than EC should be explored, which is the aim of second part of this research.

1.3 Research goals

The first part of this research aims to provide a broader study on combinations of dimension reduction techniques and clustering methods in order to facilitate the fault detection and isolation strategies. The methods that have been selected for this study are from a wide range to assure a comprehensive comparison, and it has been tried to select at least one method from different categories of clustering and dimension reduction techniques. It is demonstrated that some methods from each technique are more compatible with each other, and judicious pairing of methods will lead to more successful detection of faults present in the dataset.

The second part of this research was performed with the aim of detecting the number of clusters in an unsupervised manner. Based on the first part, the behavior of combinations of dimension reduction techniques and clustering methods were motivating to study the manner in which the structure of the data is preserved after performing different dimension reduction techniques, and this structure was studied using clustering. In other words, it is assumed that the global structure of a dataset preserves after dimensionality reduction, and comparison of various dimensionally reduced versions of the dataset and the original data can be beneficial. Thus, comparison of number of classes from clustering analysis found using different versions of data can provide insight into the original structure and cluster number. An approach is suggested for clustering method to be performed over the space of its hyperparameters to obtain a number of solutions, and a multi-objective optimization to be used on the calculated internal metrics for these clustering solutions to search for the solutions which optimize the clustering metrics, without any knowledge of the true labels. These solutions were then compared across all dimensionally reduced datasets to find the most frequent number of clusters within the solution, i.e., the number of groups which preserved their structure the most.

1.4 Thesis structure

It is noted that parts of this research is either published [32] or is submitted for publication [33] for publication as of the time of composing this thesis. Following the preceding introduction, the structure of this documented research is as follows:

- In Chapter 2 data mining methods used for this research are presented. First, dimension reduction categories, methods and examples are presented. Clustering categories and different methods are then introduced, and metrics to evaluate the performance of clusterings are then discussed.

- Chapter 3 presents the study on synergy between dimension reduction techniques and clustering methods. Previously discussed methods are tested on a case study from the benchmark Tennessee Eastman Process simulation introduced in this chapter. After the description of the approach, the results are presented and discussed at the end.

- Chapter 4 discusses the potential of combining dimension reduction techniques and clustering, with the aim of detecting the cluster number in a dataset. This chapter focuses on the methodology of the presented approach at the beginning, three case studies are introduced and the results and discussion are followed.

- Chapter 5 discusses the conclusions of this work and future directions of study for further investigations.

2 Utilized Data Mining Methods

Data mining tools play an important role in process control and monitoring systems [12]. The collected data from Distributed Control Systems (DCS) stores knowledge about the process and its behavior. If the process deviates from its normal operation, this anomaly is expected to be captured by the process measurements. Therefore, process monitoring can help avoid accidents and abnormal events in real time [12]. Furthermore, the stored data combined with data mining tools can help develop models and control strategies for further analysis of the process behavior and improve our understanding of the process. Hence, considering paths to leverage this massive data seems crucial. By allowing to study the behavior of systems through the stored data, the process can be explored in order to develop models and fault detection/diagnosis procedures and overall, improve the process operation.

Data mining tools such as clustering and dimension reduction, have proven beneficial in previous research to study faults, such as the works done by Barragan et al. [34], Thornhill et al. [35], and Gajjar et al. [36]. In general, dimension reduction techniques can be used for projection and more importantly, removing redundant and correlated data to present extracted features in lower dimensions, and enhance any following processing steps. Data clustering algorithms can be used to partition the data into groups which potentially contain data points similar to each other to identify similar classes, and using the clustering metrics on the results, the performance can be evaluated.

To use these traditional data science tools for extracting knowledge from chemical process databases, in this chapter, the methods and techniques which were used to address the research questions at hand are introduced, their approach is explained and demonstrated by examples. An example synthetic dataset has been created to demonstrate the performance of each method. Section 2.1 presents the guiding dataset which is used throughout this chapter. Section 2.2 discusses dimension reduction techniques and its categories demonstrated with examples. Section 2.3 covers clustering methods and its categories, with examples for each, and then introduces metrics to evaluate the performance of clustering outcomes.

2.1 Guiding dataset

A dataset has been created using the *make_blobs* command in Python. This function creates multiclass datasets, each class generated as a Gaussian blob. The three-dimensional dataset will be used in the subsequent sections to demonstrate the performance differences between introduced methods. The code for dataset generation and method demonstration is available in Appendix B.1. This dataset contains three classes (blobs), each containing 4 members. Two three-dimensional representations of this dataset are available in Figure 2.1. Each blob (cluster) is represented with a different color. The yellow and purple clusters

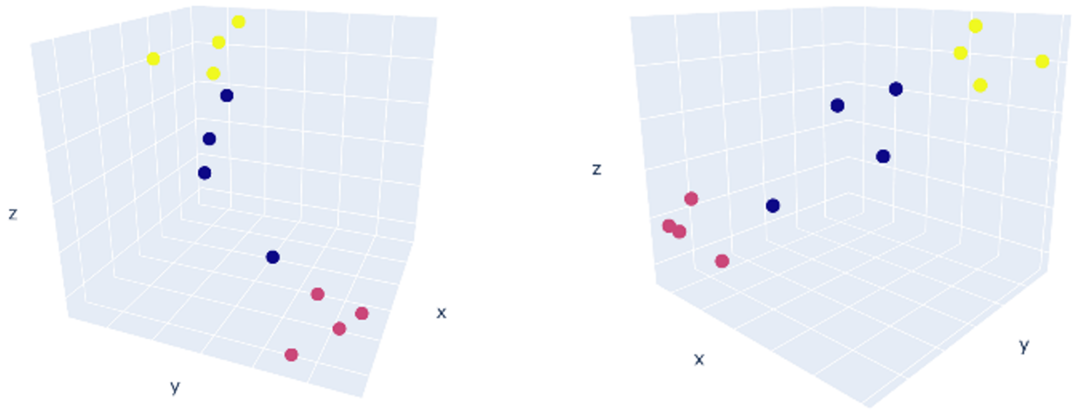


Figure 2.1: Three-dimensional projections of the guiding dataset

have a standard deviation of 0.5, and the blue cluster has a standard deviation of 1.

2.2 Dimension reduction (DR) techniques

Data acquisition from different processes resulting in large-sized datasets is not sufficient if not followed by further steps of information extraction. These datasets include multiple measurements from various instances, resulting in high-dimensional datasets. One frequently-used approach is the utilization of dimension reduction techniques. These methods belong to the category of data-visualization techniques, but are favorable compared to other methods such as parallel coordinate plots [37] and scatter plot matrices [38] because of their scalability. They can be used for information extraction and/or visualization purposes, and in the last decades, many techniques have been proposed and studied [20].

The main use of dimension reduction techniques is to remove redundant data and represent meaningful features of the raw process data in 2 or 3 dimensions (2D or 3D). Some benefits of applying dimension reduction techniques are reducing the computational complexity and avoiding the curse of dimensionality [18]. Because of these mentioned abilities, dimension reduction techniques have found a place in a variety of domains such as health sciences [39], water resource research [40] and civil engineering [41]. On top of that, while having lower number of variables facilitates data analysis, these techniques can also be used for projection purposes in 2D or 3D.

To formulate the application of DR techniques, let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ be any m-dimensional dataset, and $\mathbf{x}_i = [x_1, \ldots, x_m] \in \mathbb{R}^m$ be the ith vector of $\mathbf{X}$. Therefore, $\mathbf{X}$ is a matrix of $n \times m$. A DR technique is a function F which represents the projection $\mathbf{Y} = F(\mathbf{X})$ where $\mathbf{Y}$ is an $n \times p$ matrix, $p \leq m$. p is usually selected equal to 2 for visualization purposes, but for some methods other values can also be selected based on the study objectives. In other words, the acceptable level of retained information after the DR would be the feature deciding p. Based on the traits of DR techniques which are commonly used, three different categories have been introduced. The considered attributes leading to choose these categories are the input type of the methods, meaning whether in the high dimensions they use a distance matrix or the data points for projection, and their examined neighborhood, meaning whether local points are taken into account or they are examined globally [20]. These categories are correlation-preserving, distance-preserving and neighborhood-preserving. Details of each category and their examples are discussed next.

2.2.1 Correlation-preserving techniques

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) [42, 43] is by far the most well-known dimension reduction technique. PCA tries to extract the most information of the data by analyzing the correlation structure of the variables and keeping only the most important information [44]. The general basis of PCA assumes there exists a new set of variables, Principal Components (PCs), fewer than the original dimensions, through which the original correlated data can be expressed. These new variables are uncorrelated, and they are constructed in a way to preserve the maximum variation possible from the original data, i.e., extract information. Principal components are obtained as linear combinations of the original variables [44].

The steps of PCA for a dataset requires normalization of the dataset to zero mean and unit variance, first, due to its dependence on scale. It is noted that the components can also be found using Eigenvalue Decomposition of the original data $\mathbf{X}$. The steps to find the PCs and to project to lower dimensions are as follows:

- Calculation of the covariance matrix of the dataset

$$cov(\mathbf{X}) = \mathbf{S} = \frac{\mathbf{X}^T\mathbf{X}}{n-1} \tag{1}$$

- Eigenvalue Decomposition of the covariance matrix

$$\mathbf{V}^{-1}\mathbf{S} = \mathbf{\Lambda}\mathbf{V} \tag{2}$$

where $\mathbf{V}$ is a matrix containing the eigenvectors of matrix $\mathbf{S}$. Eigenvectors are the principal directions of the data.

Then, after sorting the eigenvalues by descending order,

- the first p eigenvectors are chosen to form a p dimensional matrix, $\mathcal{V}$ and
- using this new matrix the samples are projected onto the new space to obtain $\mathbf{Y}$, where the principal components correspond to the the columns of $\mathbf{Y}$:

$$\mathbf{Y} = \mathbf{X}\mathcal{V} \tag{3}$$

PCA and all its derivatives are considered to be correlation-preserving. The correlation between a variable and a component is an indicator of their shared information, and it is called a *loading*. Also, the percentage of the total variance which is expressed through the selected number of principal components can be calculated using the cumulative percent variance (CPV), where the variance explained by each principal component is the ratio of the variance of that PC to the total variance. Total variance in case of PCA is calculated as the trace of covariance matrix (sum of elements on the main diagonal). If all the PCs are selected, CPV will be equal to 1.

$$CPV = \frac{cov(\mathbf{Y})}{tr(\mathbf{S})} \tag{4}$$

Application of PCA on the guiding dataset introduced previously is demonstrated in Figure 2.2. The figure visualizes the second PC vs. the first PC. To further investigate the performance of PCA, as one could see in Figure 2.1, the maximum variance of the data is in the direction of 45 degrees from the x-plane starting from the purple towards the yellow cluster. Therefore, even before applying PCA, the direction of the first PC can be assessed. The next orthogonal direction which can express the next maximum variance between the data would be the direction of the second PC. So, even though points of the purple cluster can be considered close to each other in the original space, their representation in Figure 2.2 is not. The reason is that their closeness was not in the direction of the selected PCs. However, 97% of the variance of the original dataset is preserved using this representation by the first 2 PCs.

2.2.2 Distance-preserving techniques

Multidimensional Scaling (MDS)

Multidimensional Scaling [42, 45] is an example from the distance-preserving category. It is assumed that the relative distance between points in original data is more informative than their correlation, therefore the goal is to preserve that distance during the dimension reduction. Given the proximity matrix (which is simply the table of Euclidean distances), a map displaying the relative positions of the data points is constructed and is used to find the points in the lower dimensions. The general steps of the classical MDS

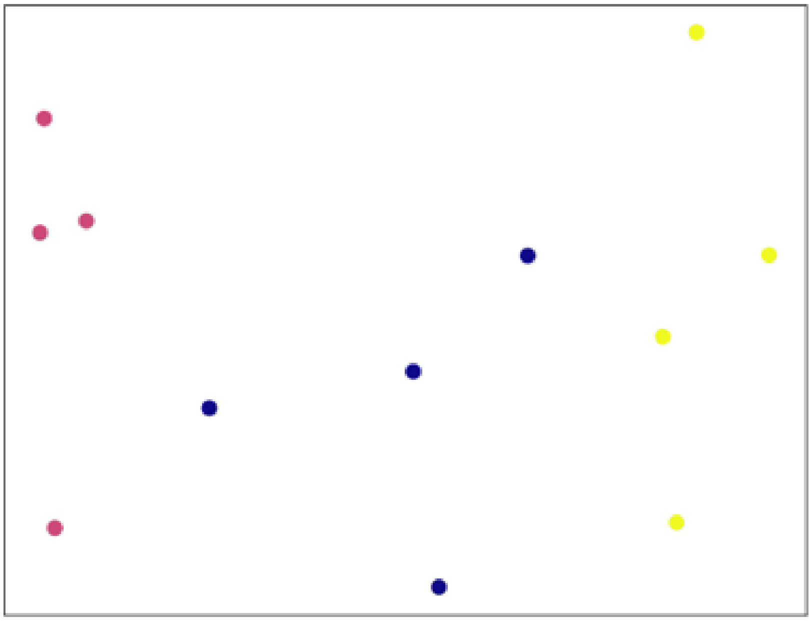


Figure 2.2: Guiding dataset - PCA reduced
PC 2 (4% variance) vs. PC 1 (93% variance)

are as follows:

- Calculation of the proximity matrix $\mathbf{D}$ using the Euclidean distance, where the (r, s) element of $\mathbf{D}$ is calculated using δ_{rs} for each pair of vectors $\mathbf{x}_r$ and $\mathbf{x}_s$ in $\mathbf{X}$:

$$\delta_{rs} = \delta_{sr} = [\sum_{t=1}^{m}(x_{rt} - x_{st})^2]^{1/2} \tag{5}$$

  Where the diagonal elements are zeros.

- Finding the new distance matrix $\mathcal{D}$ with the following steps:
  - Finding the Gram matrix which is $\mathbf{B} = \mathbf{X}^T\mathbf{X}$
  - Calculating Gram matrix from

$$\mathbf{B} = -\frac{1}{2}\mathcal{C}\mathbf{D}\mathcal{C}^T \tag{6}$$

    where $\mathbf{D}$ is the original distance matrix and $\mathcal{C}$ is the centering matrix (this matrix has the same effect as subtracting the mean of each vector from all its components), defined as following:

$$\mathcal{C} = I_n - \frac{1}{n}J_n \tag{7}$$

I_n is the identity matrix and J_n is an $n \times n$ matrix of all ones.

Since the work is done based on solely the distance, any shift/rotation of the data will produce the same proximity matrix, therefore there can be many solutions. Hence, the centering step is preformed to pin down one solution.

– $\mathcal{D}$ is found using the eigenvalue decomposition of $\mathbf{B}$, i.e., the eigenvalue matrix.

- The top p eigenvectors of the distance matrix represent the new coordinates to find $\mathbf{Y}$ where the rows of $\mathbf{Y}$ are the coordinates of points in new dimensions.

A variation of MDS is the non-metric MDS, which is more suitable for quantitative data. The algorithm has a few differences, as mentioned in the following:

- The new distance matrix $\mathcal{D}$ is assumed to be a monotonic function of the original:

  $f(\mathbf{D}) \approx \mathcal{D}$

  This monotonic transformation, $f(\mathbf{D})$, needs to be determined in order to find $\mathcal{D}$.

- The loss function (stress) is minimized by finding a new set of points

$$Stress_{\mathbf{D}} = (\frac{\Sigma(\mathbf{D} - \mathcal{D})^2}{\Sigma\mathbf{D}^2})^{1/2} \tag{8}$$

- Again, the assumption of centered configuration is applied

$$\sum_{i=1}^{n} \mathbf{x}_i = 0 \tag{9}$$

Performance of MDS on the guiding dataset is presented in Figure 2.3. Points are mapped based on their relative Euclidean distances in the original space. Therefore, it can be seen that members of the blue cluster are mostly scattered and a clear cluster structure is not detectable for the blue points.

An important matter which has been considered in the recent years is the importance of the data structure rather than the Euclidean distance. Two points may be close to each other by calculating their Euclidean distance, but they may be far in the data manifold by calculating their geodesic distance. Therefore, it is useful to also consider nonlinear manifold dimension reduction techniques to compare their results with linear methods, which is what has been mentioned up to this point.

Isometric Mapping (ISOMAP)

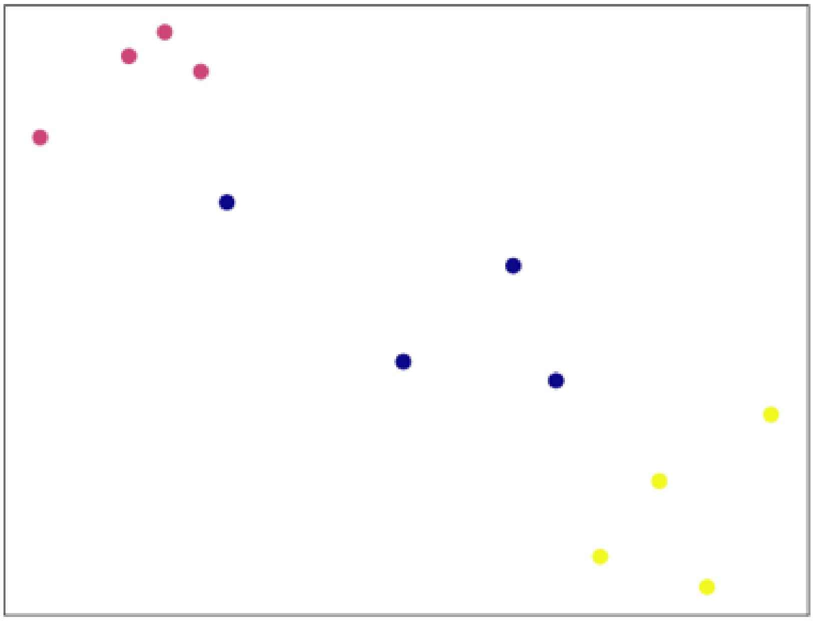


Figure 2.3: Guiding dataset - MDS reduced
Second dimension vs. first dimension

An example of the manifold learning DR techniques is ISOMAP [46], which is considered a distance-preserving technique and closely related to non-metric MDS. By assuming the inefficiency of Euclidean distance, this method tries to preserve the geodesic distance by constructing a neighborhood graph using the nearest neighbors of each point and following the steps of MDS by constructing a distance matrix. The calculation of geodesic distance depends whether two points are in the neighborhood of each other or not. The algorithm for this technique follows these steps:

- Constructing a neighborhood graph in which the neighboring points are connected. This is performed considering the nearest neighbors of each point, using a distance threshold. Instead of a distance threshold, this graph can also be created using a number for the nearest neighbors.
- Computing the shortest path between two nodes using this graph
  - Distance of two neighbors which are connected is calculated using their Euclidean distance
  - Distance of two faraway points is the sum of a series of neighbor distances, along their shortest path
- Creating the distance matrix **D** from the shortest paths
- Following the steps of classical MDS, creating a new distance matrix
- Computing the lower-dimensional embedding using the top p eigenvectors

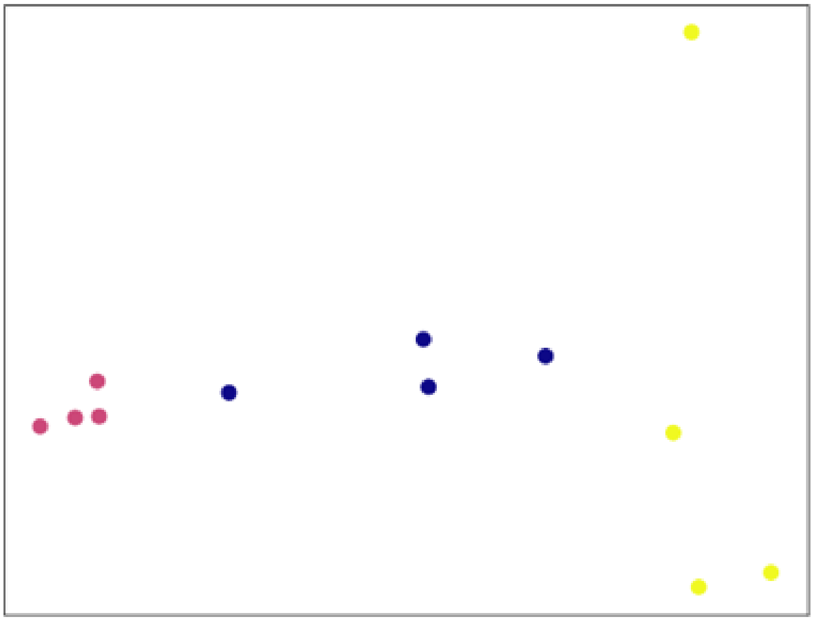


Figure 2.4: Guiding dataset - ISOMAP reduced
Second dimension vs. first dimension

As it is evident from the ISOMAP algorithm, the method's ability to detect underlying structures of a dataset is higher compared to methods which only utilize Euclidean distance. While sustaining computational efficiency, one issue with ISOMAP regards the placement of possible noise effects in the dataset. If there are outliers present which are far from the data manifold, the connectivity graph will result in a very different low dimensional embedding and can be distorted by the noise [47].

Result of dimension reduction on the guiding dataset using ISOMAP is presented in Figure 2.4. During this mapping, the blue point farthest from its cluster in the original space is close to the points of the purple cluster and is represented that way, however, the farthest yellow point from its cluster is not close to any other points of the dataset in the original space. This results in its mapping with a higher distance from other points, even members of its own cluster.

2.2.3 Neighborhood-preserving techniques

t-Stochastic Neighborhood Embedding (t-SNE)

Next category is neighborhood preserving techniques. t-Stochastic Neighborhood Embedding (t-SNE) [48] is an example of this category, which assumes local neighborhoods are more important in revealing the inner structure of the data. This method tries to preserve the local neighborhoods by creating a Gaussian probability distribution of each and projecting them to the lower dimensions using a t-student distribution while minimizing a cost function. In other words, this method tries to maximize the probability of closeness of local points in lower dimensions based on their placements in the original space. Because of the "heavy tails" probability of the t-distribution, the relative distances get more extreme, thus highlighting the neighborhoods more clearly. Since this algorithm works only with local neighborhoods, any global structure, or in other words, distance between clusters in the high dimensions gets lost in the low-dimensional representation. The method follows these general steps:

- Calculating the conditional probability of each point belonging to the neighborhood of point j using their Euclidean distance in high dimensions (meaning how likely they are to be neighbors if the distribution is Gaussian)

  $$P_{k|j} = \frac{exp(-d_{jk}/2\sigma_j^2)}{\sum\limits_{k \neq j} exp(-d_{jk}/2\sigma_j^2)} \tag{10}$$

  d_{jk} is the Euclidean distance between points x_j and its neighbors x_k, and σ_j is the standard deviation of the Gaussian distribution.

- Calculating the joint probability distribution

  $$P_{jk} = \frac{P_{k|j} + P_{j|k}}{2n} \tag{11}$$

- Building a dataset of points in a low-dimensional space

- Calculating the joint probability distribution, assuming the distribution is t-student in lower dimensions

  $$Q_{k|j} = \frac{exp(-d_{jk})}{\sum\limits_{k \neq j} exp(-d_{jk})} \tag{12}$$

- Calculating how different the distributions from high and low dimensions are from each other, and minimizing this difference using the Kullback-Leiber divergence

  $$D_{KL}(P||Q) = \sum_{\mathbf{x} \in \mathbf{X}} P(\mathbf{x}) log(\frac{P(\mathbf{x})}{Q(\mathbf{x})}) \tag{13}$$

  This expression shows how much information would be lost if some data with distribution $P(\mathbf{x})$ is expressed using distribution $Q(\mathbf{x})$, i.e., projecting from higher dimensions to lower dimensions.

Because of the non-symmetrical property of the cost function, differences between pairwise distances in lower dimensions are not treated equally. In particular, there is a large cost if highly separated points are

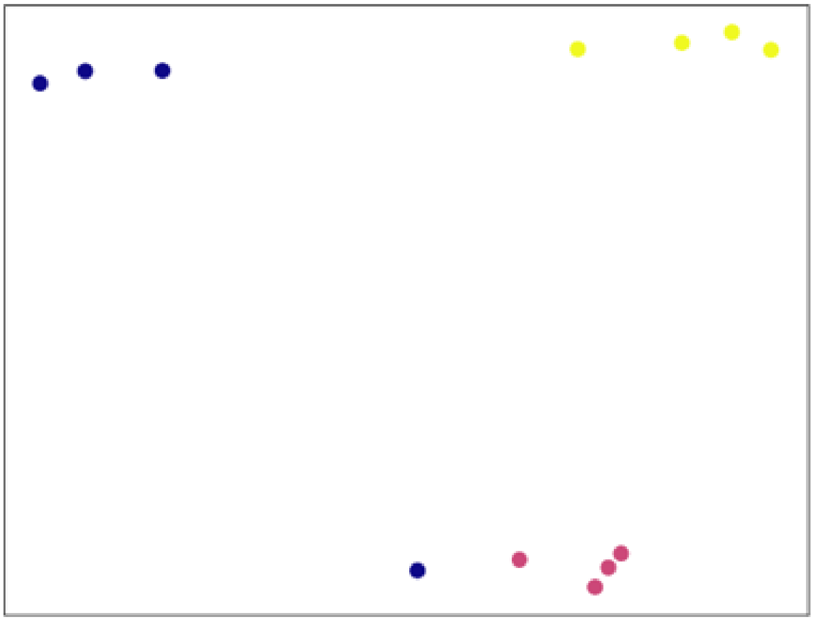


Figure 2.5: Guiding dataset - t-SNE reduced
Second dimension vs. first dimension

represented by nearby points, and there is a small cost if nearby points are represented by separated points. Therefore, most of the focus in this method is to preserve the local structure of the data.

t-SNE dimension reduction of the guiding dataset is demonstrated in Figure 2.5. In this case, features of the dataset are extracted in terms of local neighborhoods, resulting in separation of yellow points and purple points. However, the farthest blue point is closer to the purple cluster than to members of the same color and in the end it is considered in the purple neighborhood. In this representation, the distance between clusters does not have any meaning and as demonstrated, closeness of points, whether inside clusters or between clusters, is only regarding the local neighborhoods.

Uniform Manifold Approximation and Projection (UMAP)

Uniform Manifold Approximation and Projection (UMAP) [49] is in the same category of neighborhood-preserving techniques and it is similar to t-SNE, while having some differences in conditional probability calculations and cost function in order to try to preserve the global structure as well. Overall, UMAP uses manifold approximations to construct a topological representation of the high-dimensional data. Then, given some low-dimensional representation of the data, UMAP optimizes the lower dimensional data to minimize the cross-entropy between these two representations. Generally, the approach of UMAP is very similar to t-SNE, with the differences of UMAP and t-SNE as mentioned:

- Calculation of the conditional probability in high dimensions is not limited to using Euclidean distance, it utilizes exponential probability distribution, and the probabilities are not normalized. Also, the calculation of joint probability is different:

$$P_{k|j} = exp(\frac{-d_{jk} - \rho_j}{\sigma_j}) \tag{14}$$

where ρ_j for the neighborhood of point x_j is defined as following:

$$\rho_j = min\{d_{jk} | d_{jk} > 0\} \tag{15}$$

- A number for nearest neighbors is calculated to create a weighted neighborhood graph to represent the topology of the data
- The distribution in low dimensions is of the following form:

$$Q_{jk} = (1 + a(y_j - y_k)^{2b})^{-1} \tag{16}$$

  where a and b are calculated using a minimum distance parameter, and y_j and y_k are coordinates of points in the lower dimensions.

- Binary cross-entropy is minimized as the cost function between two sets of distributions:

$$CE(\mathbf{X,Y}) = \sum_j \sum_k \big[P_{jk}(\mathbf{X}) log \frac{P_{jk}(\mathbf{X})}{Q_{jk}(\mathbf{X})} + \big(1 - P_{jk}(\mathbf{X})\big) log \frac{1 - P_{jk}(\mathbf{X})}{1 - Q_{jk}(\mathbf{X})}\big] \tag{17}$$

Applying the last dimension reduction technique on the guiding dataset creates Figure 2.6. Again, by using a neighborhood-preserving technique one blue point is placed into the purple neighborhood. That being said, it is visible that blue and yellow neighborhoods are closer to each other than to purple neighborhood. This signifies the UMAP attempt to balance the local-global structure in contrast to t-SNE.

2.2.4 DR summary

It was tried to cover the most widely used DR techniques in the first part of this section. The techniques were categorized based on their approach to transform the dimensions of a dataset. Evidently, each technique has its own advantages and disadvantages, and no single technique is successful in feature extraction for all applications. Applying dimension reduction requires knowledge about the underlying approach of techniques, as well as familiarity with the dataset. Each technique can unveil different characteristics of the data, but may not be able to capture all the features.

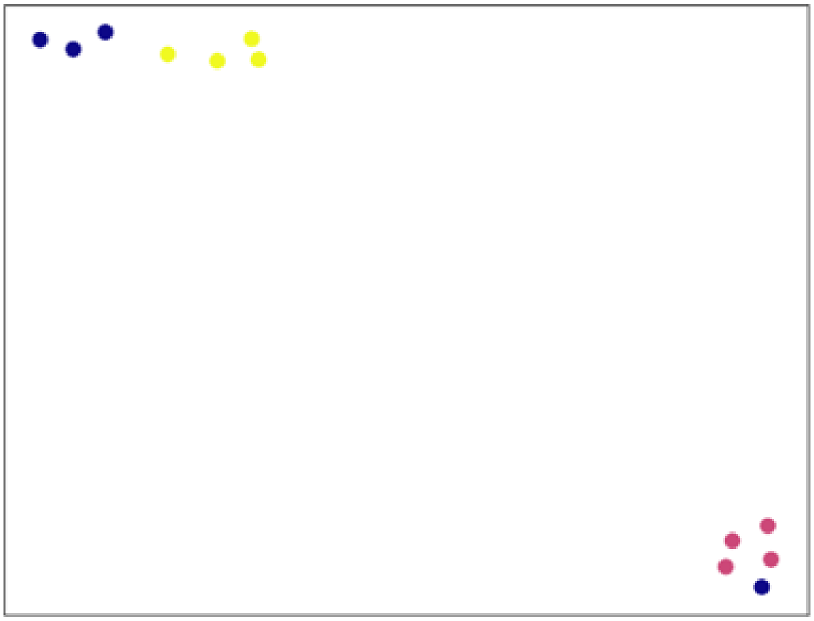


Figure 2.6: Guiding dataset - UMAP reduced
Second dimension vs. first dimension

2.3 Clustering Techniques

Clustering is one of the most important unsupervised learning methods, in which a dataset is partitioned into a number of clusters (classes) with similar characteristics, where the class label information is not known a priori. It has been widely used in many applications throughout the years and has been applied to different types of data. Some applications of clustering analysis include web search [50], biology [51], and image pattern recognition [52]. Clustering can be used to gain insight into the distribution of data and to observe characteristics of different clusters. The process of discovering groups within data can be done using various algorithms [3].

To formulate, given a dataset $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ of size $n \times m$, clustering can be formulated as a function G depending on the method, where $\mathbf{z} = G(\mathbf{X})$. $\mathbf{z}$ is a $n \times 1$ vector and $\mathbf{z} \in \mathbf{C} = [C_1, \ldots, C_{N_T}]$, where the aim of clustering is to find N_T, which is the total number of clusters and also to find the membership (assignment) of each data point. In addition, it is assumed that there exists true (or expected) set of labels defined by a $n \times 1$ vector $\mathbf{u}$, where the elements are of the subsets of true labels $\mathbf{L} = \{L_1, \ldots, L_j, \ldots, L_{N_L}\}$. These true labels can be available or unavailable in different cases. In cases where true labels are available, they can be used for conducting supervised analysis (classification), or they can be used for performance assessment of unsupervised clustering. Different clustering methods can lead to different clusterings based on their approach. Some methods are also capable of detecting outliers, and can be used for such purposes.

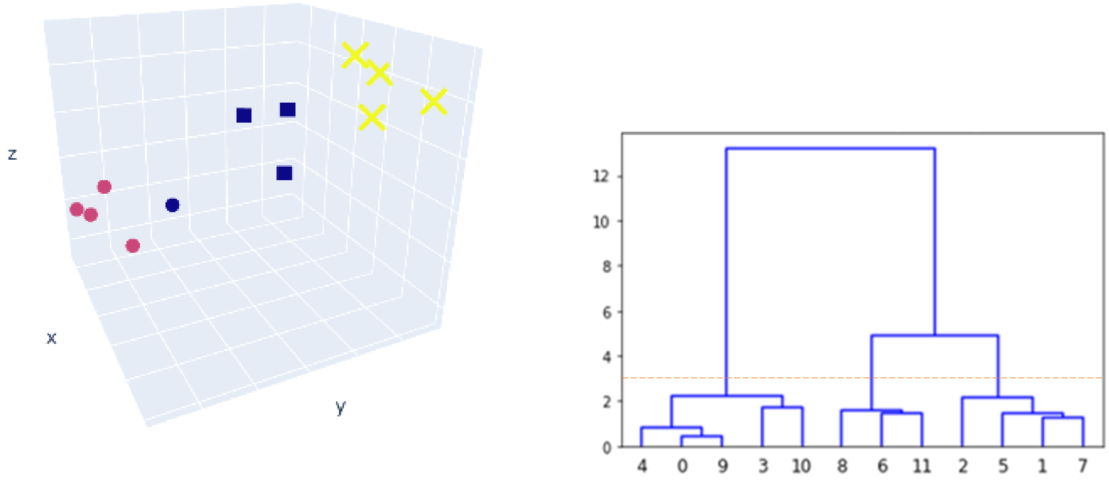
The methods discussed in this chapter partition the data into mutually exclusive clusters. Five different major categories of clustering methods have been introduced in the current section, and the categories are based on the partitioning criteria of the algorithms. These categories are connectivity-based, centroid-based, distribution-based, density-based and grid-based. After that, the metrics to evaluate the clustering performance are discussed.

2.3.1 Connectivity-based clustering

The methods in the connectivity-based category take the distance between data points into account to assign cluster memberships, where each data point is a vector. Closer data points are assumed to be in similar clusters as opposed to farther data points. The most well-known method in this category is the agglomerative hierarchical clustering [53], using a number of possible closeness measures (linkages). This method uses a bottom-up merging strategy, starting by considering every data point as a single cluster and then merging the two closest clusters together, and this merging process (fusion) is repeated until the whole dataset is one cluster. The hierarchy of these combinations is represented in a dendrogram in the shape of a tree. This tree shows how objects are grouped together. After the process is terminated, the user can choose where to cut the dendrogram, therefore defining the final number of clusters.

Assuming r and r' are two data points in $\mathbb{R}^m$, m_l is the mean for cluster C_l and $|C_l|$ is the number of objects in that cluster, five distance measures (linkages) are most widely used for this method as mentioned below. However, other metrics can also be defined to indicate the merging threshold of different clusters together.

- Minimum distance (single linkage using Euclidean distance): $= min\{d_{rr'}\}$ where $r \in C_l, r' \in C_o$
- Maximum distance (complete linkage using Euclidean distance): $= max\{d_{rr'}\}$ where $r \in C_l, r' \in C_o$
- Mean distance: $= d_{m_l m_o}$
- Average distance: $= \frac{1}{|C_l||C_o|} \sum_{1}^{|C_l|} \sum_{1}^{|C_o|} d_{rr'}$ where $r \in C_l, r' \in C_o$
- Ward's linkage: $= min\{ESS_{lo} - [ESS_l + ESS_o]\}$ where
  $ESS_l = \sum_{1}^{|C_l|} d^2_{rm_l}$ for all $r \in C_l$
  $ESS_o = \sum_{1}^{|C_o|} d^2_{r'm_o}$ for all $r' \in C_o$
  and for merged cluster C_{lo} resulting from fusion of clusters C_o and C_l:
  $ESS_{lo} = \sum_{1}^{|C_{lo}|} d^2_{r''m_{lo}}$ for all $r'' \in C_{lo}$



(a) Clusters found using hierarchical clustering

(b) Dendrogram of hierarchical clustering

Figure 2.7: Guiding dataset - application of hierarchical clustering

As a demonstration, hierachical clustering using average linkage was applied to the guiding dataset and the results are presented in Figure 2.7. In this case, each data point is assumed as a single cluster in the beginning and in each step, closest points (closeness defined using average linkage) are merged together, until there are three clusters left. For clustering demonstrations, blobs are represented with the same colors as before and their clustering assignments for each method are represented using different shapes. In Figure 2.7(a), yellow and purple points are correctly assigned to separate clusters, however, one point of the blue class is assigned to the purple cluster. Figure 2.7(b) shows the merging steps vs. the index of points. In this case, the dendrogram has been cut on three clusters at a vertical distance of about 3.

2.3.2 Centroid-based clustering

Methods in the centroid-based category find centroids in order to partition the data into a specific number of clusters. This is done by minimizing the distance of points from their closest centroid, where the centroid is the center point of the cluster. The examples chosen for this category are k-means [54] and k-medoids [55]. The main difference between these two methods is their selection of the centroid. For k-means, the algorithm uses the number of clusters, N, as the input by the user, and tries to separate the data into N groups of equal variance, while minimizing the Sum of Squared Errors (distances) for members of a cluster from the centroid of that cluster (e_l):

$$SSE = \sum_{l=1}^{N} \sum_{r \in C_l} dist(r, e_l)^2 \tag{18}$$

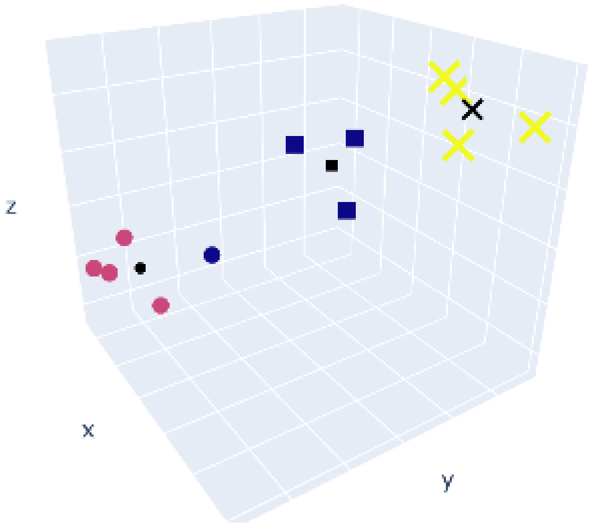


Figure 2.8: Guiding dataset - application of k-means clustering

In other words, the distance from the points in the cluster to their centroid is squared, and the distances are summed. This objective function tries to detect N clusters as compact and as separate as possible. The centroids are randomly chosen in the beginning and get updated through each iteration of the algorithm to be the mean of their cluster members (iterative relocation technique). This randomness slightly affects the results in every run. The k-means algorithm follows these steps:

- Selection of N centroids randomly
- and, until no change, repeating these steps:
  - assigning each point to its closest centroid
  - updating the centroid to be the mean of the cluster members

Clusters detected in the guiding dataset using k-means are presented in Figure 2.8. Each cluster centroid is represented in black using the same shape of its cluster members. Again in this case, the farthest blue point has been assigned to the same cluster as the purple points, all presented with a circle.

k-medoids is similar to k-means, but instead minimizes a sum of general pairwise dissimilarities, and the medoids (centroids) are from the points in the dataset. k-medoids is more robust than k-means when outliers are present in the data, because a medoid is less influenced by outliers than a mean [3]. k-medoids algorithm includes the following steps:

- Random selection of N initial representatives from the objects in the data

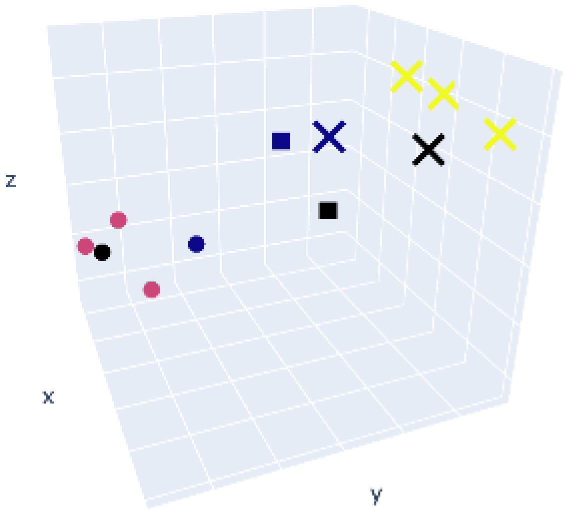


Figure 2.9: Guiding dataset - application of k-medoids clustering

- and, until no change, repeating these steps:
  - assigning each point to its closest representative
  - computing the cost of swapping the representative object with a random nonrepresentative object
  - If the cost is negative, updating the representative with the nonrepresentative to be form a new set

Figure 2.9 demonstrates the performance of k-medoids. It is visible that the centroids in this case are from the data points (marked in black). To point out another difference from k-means, using this method, one blue point is clustered with the yellow point, since it is closer to the x cluster centroid than to the square cluster centroid.

2.3.3 Distribution-based clustering

The main idea exploited in the third category, distribution-based clustering, is the assumption that members of a cluster most likely belong to the same distribution. In other words, different categories of data form latent distributions. Points are assigned to clusters based on their probability of belonging to a distribution. Gaussian mixture model (GMM) [56] is an example that assumes that the data is constructed of multiple Gaussian distributions. The method follows an algorithm as described below:

Assume there are N Gaussian distributions, each with a center, μ_l, and standard deviation, σ_l, such that

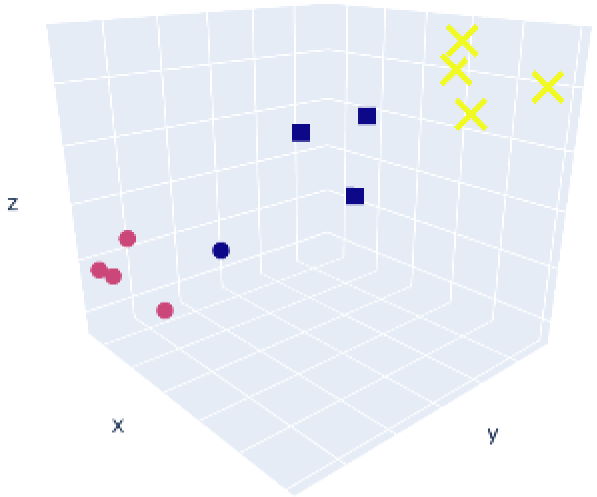


Figure 2.10: Guiding dataset - application of GMM clustering

$\Theta_l = (\mu_l, \sigma_l)$ and $\Theta = [\Theta_1, \ldots, \Theta_N]$. For the dataset $\mathbf{X}$ we have:

$$P(x_i|\Theta_l) = \frac{1}{\sqrt{2\pi\sigma_l}} e^{-\frac{(x_i-\mu_l)^2}{2\sigma^2}} \tag{19}$$

Assuming each cluster has the same probability $\omega_1 = \omega_2 = \cdots = \omega_N = \frac{1}{N}$, where each probability is an indicator of some instance is sampled from that cluster, Eq. (19) can be written as:

$$P(\mathbf{X}|\Theta) = \frac{1}{N} \prod_{i=1}^{|\mathbf{X}|} \sum_{l-1}^{N} \frac{1}{\sqrt{2\pi\sigma_l}} e^{-\frac{(x_l-\mu_l)^2}{2\sigma^2}} \tag{20}$$

Then, Eq. (20) is maximized using the Expectation-Maximization (EM) algorithm to find the parameters of each Gaussian to finally find the probabilities of each sample belonging to any Gaussian, where $|\mathbf{X}|$ is the number of all elements in the dataset. In general, EM algorithm tries to estimate the parameters in a statistical model a posteriori. In this algorithm, a set of initial parameters are selected and the algorithm iterates until the clustering cannot be improved (i.e., the results converge). This is also similar to the k-means algorithm. Each iteration of EM has two steps [3]:

1. The Expectation step, where objects are assigned to clusters based on the current parameters.
2. The Maximization step where new parameter are searched for based on the expected likelihood in the probabilistic model.

Clusters found using GMM are presented in Figure 2.10. Based on three Gaussian distributions, yellow

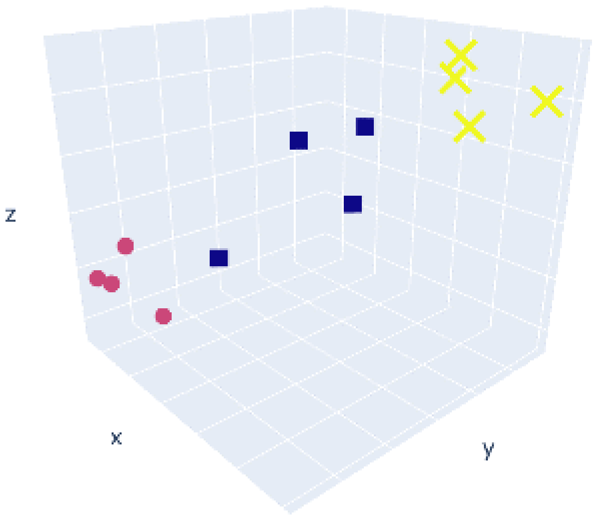


Figure 2.11: Guiding dataset - application of DBSCAN clustering

points are assigned to a single cluster, purple points with the closest blue point are assigned to another cluster and the rest of the blue points are members of another cluster.

2.3.4 Density-based clustering

In the fourth category, which includes density-based methods, clusters are defined as areas of high density separated by areas of low density. Higher density is defined as smaller regions with higher number of samples, and based on this definition, all methods in this category are able to perform outlier detection and can work with clusters which are not convex.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [57] is one example of this category. This method finds core samples as data points which have a minimum number of samples *MinPts* within a specified distance *eps* around them, and therefore finding dense areas using that distance, where the dense regions are clusters. DBSCAN algorithm is described below:

Starting with an *unvisited* object r, and repeating until are objects are marked as visited:

- if there are at least *MinPts* within a *eps* distance of r, a new cluster is created with r as a member
- for each unvisited point r' in the neighborhood of r, if there are *MinPts* within a *eps* distance, add r' to the created cluster and add combine its neighbors with neighbors of r
- if the density around r does not meet the threshold, mark r as noise or an outlier

Clusters found using DBSCAN are presented in Figure 2.11. Using a proper density threshold with *MinPts* and *eps* hyperparameters, the method is able to correctly detect clusters and assign points, while

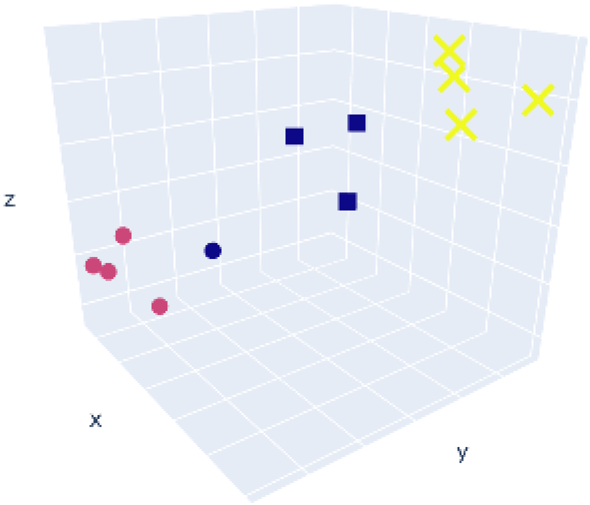


Figure 2.12: Guiding dataset - application of OPTICS clustering

this correct clustering was not possible with any of the methods introduced till now. This demonstrates the potential of density-based clustering methods. However, selection of the hyperparameters is not straightforward.

Ordering Points To Identify the Clustering Structure (OPTICS) [58] is very similar to DBSCAN, while trying to overcome the difficulty of parameter selection. DBSCAN parameters are not very intuitive to select, but the results of the clustering is highly dependent on them. This method has a major difference from DBSCAN: OPTICS orders the points to prioritize the memberships, so that objects are sorted by their reachability distance of their respective closest core samples. The reachability-distance of an object is the smallest distance from a core object. Similar to algorithm of DBSCAN, points are visited and marked as core samples (or outliers) based on their neighborhood density; Meanwhile, points in a neighborhood are sorted based on their reachability distance as well. This relaxes the *eps* threshold and allows for variable density cluster identification in a single dataset using a *min_samples* parameter.

OPTICS clustering relaxes the *eps* by setting it to infinity as a default, and by adjusting only the *min_samples* using integer values the most accurate clustering assignment is presented in Figure 2.12. In this case, one blue point is again assigned to the same cluster as purple points, and other points are correctly clustered.

Hierarchical density-based clustering (HDBSCAN) [59] also works similar to DBSCAN, but it has a hierarchical approach. The clusters are searched for using a range of distances, but instead of the size of the region determining this separation cut, the cut is placed where the number of small clusters and outliers are reasonable, i.e., the formed clusters are more stable. The input parameter is the *minimum_cluster_size* to help the algorithm make the cut in the dendrogram. The algorithm for this method starts with transforming the space according to the density/sparsity based on the populations, then building the minimum spanning tree to connect or merge dense areas. Dense areas are detected using the *minimum_cluster_size* and connected to each other one by one in a hierarchy to create a tree. This is performed using a distance weighted graph, which assigns more weight to closer dense areas, and less weight to clusters which are far from each other. In the end, the algorithm extracts the stable clusters from the condensed tree, by cutting the edges which have a small weight.

Best performance achieved by HDBSCAN is presented in Figure 2.13. This method separated the dataset into two clusters, splitting the blue points between the other two clusters. Using a *minimum_cluster_size* of 2, top 3 blue points are merged with the yellow cluster and the bottom blue point is assigned to the purple cluster.

2.3.5 Grid-based clustering

Methods in the last category divide the space into finite number of cells to form a grid structure and calculate the density of each grid to find the clusters. Therefore, they can be viewed as space-driven methods, as opposed to the methods discussed so far which are data-driven. These methods are called grid-based clustering methods [3].

An example is CLustering In QUEst (CLIQUE) [60]. This method partitions high-dimensional data into

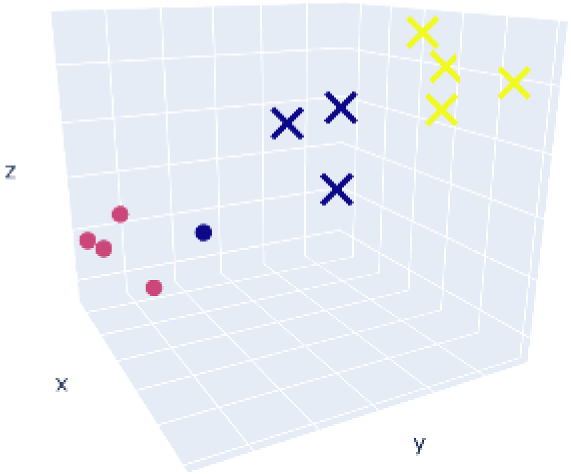


Figure 2.13: Guiding dataset - application of HDBSCAN clustering

non-overlapping subspaces and uses these subspaces to find clusters. It tries to identify dense cells where their number of objects exceeds the density threshold (d_{th}), and sparse cells which are less likely to contain clusters. This identification is done by based on d_{th} and *interval*, which determines the partitioning of each dimension. After that, the method iteratively joins two dense cells if they share the same intervals and dimensions in the space. This is repeated until a cluster cannot be further extended in any dimension. CLIQUE has a greedy approach, in which it starts with an arbitrary dense cell and then tries to extend it to a maximal region covering more cells. It should be noted that the selection of the intervals, i.e., grids, is particularly important in this algorithm in order to extract stable clusters. Also, if the clusters have varying density, the selection of the threshold is rather influential in the final detection of shapes [3].

Figure 2.14 shows the results of CLIQUE clustering using an interval of 5 and each dimension of the original space has been separated into 5 intervals. Threshold is used to determine the minimum number of points a cell can contain to be considered a dense cell. In this case, the threshold has been set to zero, meaning any cell which contains as minimum as one point can be a part of a cluster. As demonstrated, this method has detected four clusters in total, successfully assigning yellow and purple points and splitting the blue points into two different clusters. This indicates that an interval of 5 does not merge the cell containing the farthest blue point with any other cells and their points.

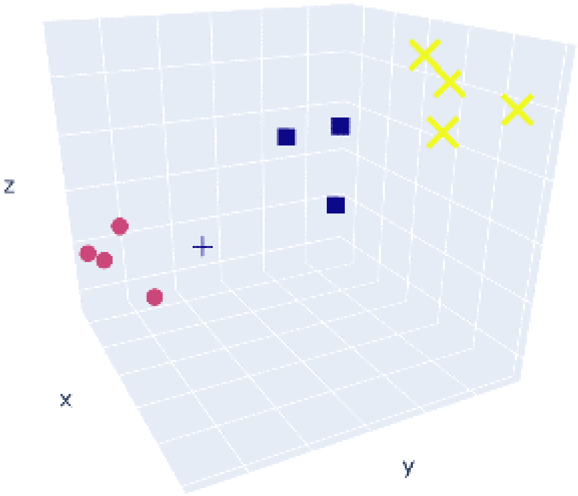


Figure 2.14: Guiding dataset - application of CLIQUE clustering

2.3.6 Performance evaluation metrics

After performing clustering, it is important to assess the labelling generated by the analysis, and in some cases to compare the performance of different clustering methods. Consider two assignments of the same vector $\mathbf{x} = [x_1, x_2, \ldots, x_n]$, $\mathbf{C} = [C_1, \ldots, C_l, \ldots, C_{n_C}]$ with n_C clusters and $\mathbf{L} = [L_1, \ldots, L_o, \ldots, L_{n_L}]$ with n_L clusters. Performance assessment of clustering methods depends on an important piece of information about the data; whether the true assignments of data points are known or not. If true labels are known, the clustering assignments ($\mathbf{C}$) are compared with these true labels ($\mathbf{L}$). Cases with known true labels are usually used for training and classification purposes. There are a number of important metrics which are calculated using the true labels and clustering assignments, and are called external metrics [3]. The ones introduced in this research are AMI, ARI and V-measure. For the objectives of this research, these external indices are only used for performance assessment and not for supervised learning. Obviously, these indices provide a more accurate assessment of how the clustering was performed.

Adjusted Mutual Information (AMI)

AMI [61] is the adjustment of Mutual Information score to measure the agreement between true labels and assignments while accounting for chance, and is an *external* metric. The mathematical description follows the steps below:

Considering $\mathbf{C}$ and $\mathbf{L}$ and selecting an object randomly from $\mathbf{x}$, the probability of the object falling into cluster C_l is expressed as:

$$P(l) = \frac{|C_l|}{n} \tag{21}$$

where $|C_l|$ is the number of elements in cluster C_l. The entropy ϕ of a set represents the amount of uncertainty for a partition set:

$$\phi(\mathbf{C}) = -\sum_{l=1}^{n_C} P(l) log(P(l)) \tag{22}$$

$$\phi(\mathbf{L}) = -\sum_{o=1}^{n_L} P(o) log(P(o)) \tag{23}$$

The mutual information between the sets $\mathbf{C}$ and $\mathbf{L}$ is then expressed as:

$$MI(\mathbf{C}, \mathbf{L}) = \sum_{l=1}^{n_C} \sum_{o=1}^{n_L} P'(l,o) log(\frac{P'(l,o)}{P(l)P(o)}) \tag{24}$$

where $P'(l,o)$ is the probability that a sample point belongs to cluster C_l in the assignment $\mathbf{C}$ and to cluster L_o in the assignment $\mathbf{L}$. The expected value of MI is calculated according to the following:

$$E[MI(\mathbf{C},\mathbf{L})] = \sum_{l=1}^{n_C}\sum_{o=1}^{n_L}\sum_{n_{lo}=(a_l+b_o-n)^+}^{min(a_l,b_o)} \frac{n_{lo}}{n} log(\frac{n \cdot n_{lo}}{a_l b_o}) \frac{a_l!b_o!(n-a_l)!(n-b_o)!}{n!n_{lo}!(a_l-n_{lo})!(b_o-n_{lo})!(n-a_l-b_o+n_{lo})!} \tag{25}$$

For an easier notation, $a_l = |C_l|$ is the number of elements in C_l and $b_o = |L_o|$ is the number of elements in L_o. $n_{lo} = |C_l \cap L_o|$ denotes the number of objects common to clusters C_l and L_o. The variables a_l and b_o are the partial sums of the contingency table of predicted and true (expected) labels: $a_l = \sum_{o=1}^{n_C} = n_{lo}$ and $b_o = \sum_{l=1}^{n_L} = n_{lo}$. $(a_l + b_o - n)^+$ denotes $max(1, a_l + b_o - n))$. The final AMI score is calculated as

$$AMI = \frac{MI - E[MI]}{mean(\phi(\mathbf{C}), \phi(\mathbf{L})) - E[MI]} \tag{26}$$

The AMI score is bounded between 0 and 1, where assuming that one of the assignments contains the true cluster labels, the higher values correspond to a better performing clustering assignments.

Adjusted Rand Index (ARI)

ARI [62] is another external metric, which measures the similarity between clustering assignments and true labels. This score is bounded between 0 and 1, and 1 is the perfect match score.

Given $\mathbf{C}$ and $\mathbf{L}$, their similarity can be represented in a contingency table where each entry n_{lo} shows the number of samples in common between each C_l and L_o as in Table 2.1.

Table 2.1: Contingency table of ARI calculation

$\mathbf{C} \setminus \mathbf{L}$	L_1	L_2	$\ldots$	L_{n_L}	sums
C_1	n_{11}	n_{12}	$\ldots$	n_{1n_L}	a_1
C_2	n_{21}	n_{22}	$\ldots$	n_{2n_L}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
C_{n_C}	n_{n_C1}	n_{n_C2}	$\ldots$	$n_{n_Cn_L}$	a_{n_C}
sums	b_1	b_2	$\ldots$	b_{n_L}	

And ARI is calculated using the permutation method as:

$$ARI = \frac{\sum_{lo}\binom{n_{lo}}{2} - \Big[\sum_l \binom{a_l}{2} \sum_o \binom{b_o}{2}\Big] / \binom{n}{2}}{\frac{1}{2}\Big[\sum_l \binom{a_l}{2} + \sum_o \binom{b_o}{2}\Big] - \Big[\sum_l \binom{a_l}{2} \sum_o \binom{b_o}{2}\Big] / \binom{n}{2}} \tag{27}$$

V-measure

Another external metric is calculated using two other performance metrics: Homogeneity (h) and Completeness (c). The former is calculated based on whether each cluster only contains members of a single class and the latter regards whether all members of a given class are assigned to the same cluster. Each are calculated as follows:

$$h = 1 - \frac{\phi(\mathbf{C}|\mathbf{L})}{\phi(\mathbf{C})} \tag{28}$$

$$c = 1 - \frac{\phi(\mathbf{L}|\mathbf{C})}{\phi(\mathbf{L})} \tag{29}$$

where $\phi(\mathbf{C})$ and $\phi(\mathbf{L})$ are the entropy values of sets $\mathbf{C}$ and $\mathbf{L}$, while $\phi(\mathbf{C}|\mathbf{L})$ is the conditional entropy of the classes, given the cluster assignment:

$$\phi(\mathbf{C}|\mathbf{L}) = -\sum_{l=1}^{|\mathbf{C}|}\sum_{o=1}^{|\mathbf{L}|} \frac{a_{o,\mathbf{L}}}{N} \cdot log(\frac{a_{l,\mathbf{L}}}{b_o}) \tag{30}$$

Here $a_{l,\mathbf{L}}$ is the number samples belonging to class $\mathbf{C}$ and assigned to class $\mathbf{L}$. The harmonic mean of these scores is called the V-measure [63], bounded between 0 and 1, where higher values represent more accurate clustering assignments:

$$v = 2 \cdot \frac{h \cdot c}{h + c} \tag{31}$$

As most cases have to deal with unsupervised situations, in other words, the true labels of the data are not known, the performance assessment of the clustering is done using the *internal* features of the found clusters. These features can be placed into two categories: cohesion measures and separation measures [64]. Cohesion can be interpreted as the tightness of the found clusters, and separation is how “far” the clusters are from each other. Clusters are ideally defined as groups of points which have high cohesion and high separation. Although, the tightness of each of clusters and distance of pair of clusters are the same metrics that the clustering methods optimize to detect clusters. Therefore, these evaluation metrics and most clustering methods confirm each other’s results, hence the results might not correctly represent the actual clusters and structure of the data. Next, *internal* metrics are introduced.

Silhouette Coefficient

The most well-known internal metric is the Silhouette coefficient [65]. Similar to most of the internal validation metrics, this metric is the ratio of cohesion to separation, and maximization of the metric is favorable. For a data point r in cluster C_l, the simplified formula is:

$$sc(r) = \frac{B - A}{max(A, B)} \tag{32}$$

where A is the mean distance between a sample and all other points in the same cluster C_l:

$$A(r) = \frac{1}{|C_l| - 1} \sum_{k \in C_l, r \neq k} d(r, k) \tag{33}$$

and B is the smallest mean distance between a sample and all other points in the next nearest cluster C_o:

$$B(r) = \min_{o \neq l} \frac{1}{|C_o|} \sum_{k \in C_o} d(r, k) \tag{34}$$

For this metric to be maximized (it has an upper bound of 1 and a lower bound of -1), it is ideal for A to be much smaller than B. A high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. The Silhouette coefficient, sc, is calculated for all points and the overall sc is the mean over all data points.

Davies-Bouldin (DB) Index

Another example of internal metrics is the Davies-Bouldin (DB) index [66], which compares the distance between clusters with the size of the clusters themselves. Two scores are calculated, T and M. T is the average distance between each point of cluster and the centroid of that cluster (cluster diameter).

$$T_l = \left(\frac{1}{|C_l|} \sum_{j=1}^{|C_l|} |\mathbf{x}_j - e_l|^p \right)^{\frac{1}{p}} \tag{35}$$

where $\mathbf{x}_j$ is the feature vector assigned to the cluster, e_l is the centroid of the cluster, $|C_l|$ is the size of the cluster and p is usually 2 to consider a Euclidean distance (2-norm). M is the distance between cluster centroids:

$$M_{l,o} = ||e_l - e_o||_p \tag{36}$$

A value of R_{ij} is then calculated for each pair of clusters:

$$R_{lo} = \frac{T_l + T_o}{M_{l,o}} \tag{37}$$

The DB index is found using the maximum of R_{lo} values:

$$DB = \frac{1}{k} \sum_{l=1}^{k} \max_{l \neq o} R_{lo} \tag{38}$$

where k is the number of clusters.

Based on previous definitions, it would be ideal to have a high value of M and a low value of T, therefore a low value of DB index. The lower values would then indicate a model with better separation between clusters. The score has a minimum of zero.

Also, there are some other examples such as Dunn index [67] and Calisnki-Harabasz index [68] with similar structures. Clearly, a problem arises in cases that do not have convex clusters, including cases with elongated or arbitrary-shaped clusters, as these clusters will not achieve good scores regarding these metrics. Also, most of these metrics do not account for any possible unclustered data points (outliers) which might have been found using some of the clustering methods. Hence, a need for an internal clustering evaluation metric with different features becomes necessary.

Density-Based Clustering Validation (DBCV) Index

There have been some developments in proposing new metrics which work with other features of the clusters, thus could be more suitable to assess the performance of density-based clustering methods. The most well-developed one is Density-Based Clustering Validation Index (DBCV) [69].

Similar to the objectives of density-based clustering methods, this metric considers the relative density connections between pairs of objects. The all-points-core-distance, which is the inverse of the density of each object with respect to all other objects inside its cluster, is calculated using the following formula:

$$a_{pts}coredist(r) = \left(\frac{\sum_{k=2}^{n_k} \left(\frac{1}{KNN(r,k)} \right)^d}{n_k - 1} \right)^{-\frac{1}{d}} \tag{39}$$

This value is calculated for each point. $KNN(r,k)$ is the distance between object r and its kth nearest neighbors, in this case all other objects in its cluster, and therefore $1/KNN$ could be interpreted as a density measurement. Then, for all pairs of r and r' objects in the cluster, the Mutual Reachability Distance (MRD) is found as:

$$d_{mreach}(r, r') = max\{a_{pts}coredist(r), a_{pts}coredist(r'), d(r, r')\} \tag{40}$$

Based on this calculation, dense points (with low core distance) remain with the same distance and sparser points are moved further away to be at least one core distance away form each other. Then, a Mutual Reachability Distance Graph is created, which is a complete graph with objects in the dataset as vertices and the MRD between pairs as the weight of each edge. From the graph, a Minimum Spanning Tree (MST) is built to decide where and how clusters should be defined. The tree is built one edge at a time, starting and moving forward with the lowest weight edges that connect the tree to a disconnected vertex, such that in the end, there is no disconnection of components. This process is repeated for all clusters.

Two features are defined here: the Density Sparseness of a cluster, DSC, defined as the maximum edge of its corresponding MST, and the Density Separation of a Pair of Clusters, $DSPC$, defined as the minimum MRD (d_{mreach}) between objects of a cluster and objects from the other clusters. Both can be interpreted as density-equivalents of cohesion and separation which was mentioned previously. Then, the Validity Index of a cluster is calculated as follows:

$$V_C(C_l) = \frac{\min_{1 \leq o \leq j, o \neq l} \big(DSPC(C_l, C_o) \big) - DSC(C_l)}{max\Bigg(\min_{1 \leq o \leq j, o \neq l} \big(DSPC(C_l, C_o), DSC(C_l) \big) \Bigg)} \tag{41}$$

And the DBCV index is found using the weighted average of the Validity Index of all clusters:

$$DBCV = \sum_{l=1}^{i=k} \frac{|C_l|}{N} V_C(C_l) \tag{42}$$

where $|C_l|$ is the size of a cluster and N is the total number of objects under evaluation. This score is bounded between -1 and +1, where negative values are cases when density inside a cluster is lower than the between-cluster density and greater values indicate better solutions.

Overall, since any clustering method has a set of parameters to adjust in order to run, these metrics can also be used for hyperparameter tuning of clustering algorithms as well.

3 Studying the Synergy between Dimension Reduction and Clustering Methods to Facilitate Fault Classification

3.1 Introduction

Detection of abnormal behaviors (faults) in chemical process systems is a challenging task due to the vast amount of data collected through measurements associated with distributed sensor networks. In these multivariate systems, gaining insight into the inner structure of collected high-dimensional data is essential for exploring and discovering paths to assist big data technologies. This exploration is facilitated by the use of many tools, such as dimension reduction techniques and clustering methods. Many fault detection methods utilizing high-dimensional historical data have been developed throughout the years and benefit from the use of dimension reduction techniques and clustering methods, or both. Each of these tools have separately been studied in the context of various applications and evaluated based on their success in class representation [20, 70]. It should be pointed out that the attributes of dimension reduction and clustering methods have a direct effect on the outcomes. Therefore, a study on such attributes will help discovering how dimension reduction techniques and a common unsupervised learning tool, clustering, can complement each other in the field of fault detection. Utilizing combinations which complement each other improves information extraction from the data and the fault detection process. There have been previous studies such as the work in [18] demonstrating some methods are more compatible with each other. The reason of such observations is explored and discussed in the current chapter by studying a wider range of methods.

In the present chapter, to help discovering methods which complement each other in the field of fault detection and to demonstrate the performance of a fault detection strategy, the synergy between dimension reduction techniques and clustering methods is studied. As shown in Figure 3.1, the proposed approach consists of a first step where the data containing different states are normalized, followed by a dimension

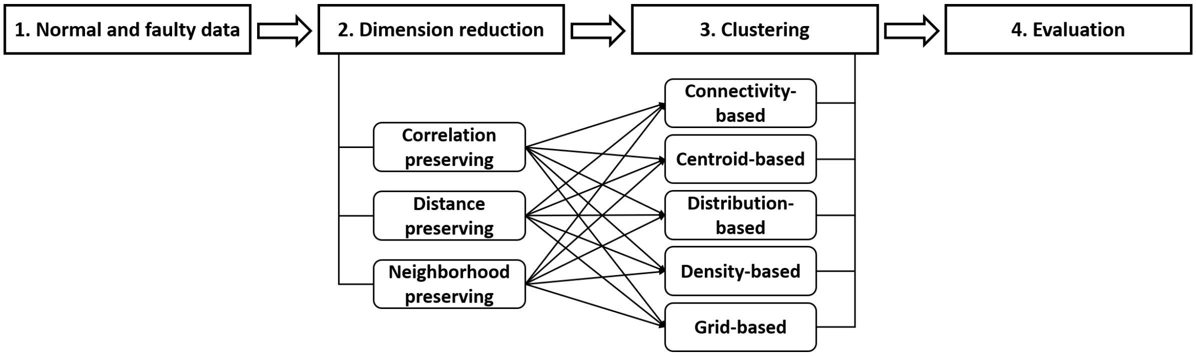


Figure 3.1: A step-wise summary to study the synergy of DR and clustering methods

reduction step then a clustering step, ending with the evaluation of the results. Different combinations of categories are tested and characteristics of each group of methods is taken into account to match each class of dimension reduction technique with one or more groups of clustering methods to achieve the best overall functionality of the fault detection process. A case study has been carried out on a dataset containing three different types of faults from the Tennessee Eastman Process simulator. The outcomes can be substantially improved by judiciously pairing the dimension reduction approach with a method from the appropriate group of clustering methods.

Tennessee Eastman Simulator and its dataset, the proposed approach and the discussion of the results are presented next.

3.2 Case study: Tennessee Eastman Process

Process monitoring systems are used to identify the operating states of a plant through a large number of sensors. Extensive amounts of variables are measured to detect and manage faults in real-time, most-importantly for safety purposes. The collected data from these sensors are useful for data-driven approaches to develop and implement effective fault detection and classifications. In order to compare the performance of different data-driven fault detection methods, a benchmark is required.

The Tennessee Eastman Process (TEP) is a popular benchmark for fault detection and monitoring studies. It is a simulation of a multimode continuous chemical process plant which was first introduced in 1993 [71], and has played an important role in the area of control design and process monitoring. A revised model of the process simulator implemented in MATLAB was introduced in 2015 [72] which included improvements compared to previous models. The datasets utilized for this research were generated using this simulator.

TEP consists of five major units (a reactor, a product condenser, a recycle compressor, a vapor-liquid separator, and a product stripper) and eight components in total. Four reactants A, C, D, E, and inert B are fed to the reactor to produce two products G and H and a by-product F. The reactor product stream is then cooled through a condenser and then fed to the separator. The vapor from the separator is recycled to the reactor through the compressor, and the liquid from the separator is transferred to the stripper. The original simulation has 41 measured variables and 12 manipulated variables, 53 variables in total, as well as 20 predefined faults of different types that can be introduced to the system, such as step changes, random variations and slow drift in reaction kinetics. The process flow of this chemical plant is presented in Figure 3.2. Details of the system variables and system faults are provided in Appendix A.

From the revised simulator, three different faults of varying types were selected to create a dataset to be utilized in different parts of this study. The dataset was generated specifically to study two systems, (i) containing multiple states, meaning faults as well as normal operation, and (ii) containing different types of faults. The selected faults were Fault 2 which is a step change in component B composition in stream 4, Fault 13, a slow drift in the reaction kinetics taking place in the reactor, and Fault 14, the sticking of the reactor cooling water valve. In order to assess and study the synergy of different methods, faults were selected from the ones which are simpler to detect. Each of these faults were separately active for a constant 20 minutes, and then turned off right before the next period started. The dataset also contains 20 minutes of normal operation (without any faults). For all periods, the sampling time is 1/3 second, which corresponds to a frequency of 180 samples a minute. The storage sampling frequency (which creates the historical database) is 1 minute, which corresponds to a frequency of 1 sample in one minute. Time constant is 1 at all times for all changes. This high-dimensional dataset contains four states in total, without transitional states between different operations. All 53 variables where initially included in the dataset before any preprocessing. It should be noted that selection of different sets of variables for studies such as clustering can lead to different clustering results. Therefore, no variables were omitted during dataset generation.

3.3 Proposed approach

In order to study the synergy between categories of dimension reduction techniques and clustering methods, the methods mentioned in Chapter 2 were tested on the dataset generated using the Tennessee Eastman

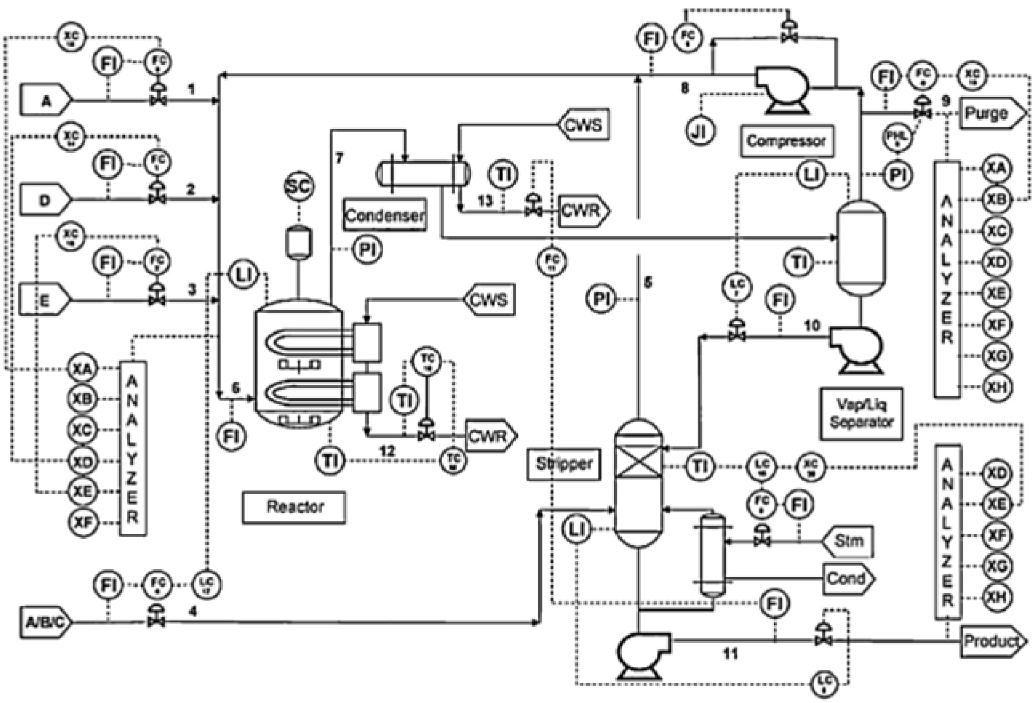


Figure 3.2: Process flow of the Tennessee Eastman Process

Process (TEP) simulator introduced previously. After the data generation, the rest of the steps are carried out using Python programming language, and the code is available in Appendix B.2, mostly using the scikit-learn package [73]. Dimension reduction techniques in this work include all examples from correlation-preserving methods, distance-preserving methods and neighborhood-preserving methods, the 5 techniques mentioned previously. The categories of clustering methods are the connectivity-based clustering (agglomerative hierarchical), centroid-based clustering (k-means and k-medoids), distribution-based clustering (GMM), density-based clustering (DBSCAN and OPTICS) and grid-based clustering (CLIQUE), 7 methods in total.

The generated dataset consisting of the three faults and one normal operational state (without any faults) is first preprocessed by removing the constant variables which didn't change in time throughout the process, and normalizing the rest of dataset to a zero mean and unit variance . A dimension reduction technique is applied to the dataset followed by a clustering method. In this manner, every combination of dimension reduction techniques and clustering methods were tested to find the categories that complement each other more successfully in isolating the different faults. All the dimension reduction techniques were used to obtain two-dimensional data, and the parameters for each clustering method were searched for based on the external metrics described in Chapter 2. They were selected such that the clustering would result in the highest possible scores for maximum number of metrics; In other words, the clustering assignments would match the true labels of the four different states as much as possible. Hence, better scores (closer to 1) do not necessarily correspond to correct number of clusters, but they may contain some extra clusters or some combined clusters. This study is only intending to discover the complementary methods, and it is not researching the abilities of each individual method. Hence, the study is supervised, meaning that true labelings of dataset were used to maximize the ability of clustering methods in detecting different states. This supervised clustering is slightly different from the definition of classification in a detail: the clustering is not using training data to predict new data, meaning all data is used at once to detect classes. Whereas in classification, new data is classified into defined groups based on previous training.

3.4 Results and discussion

The ability of a clustering method to accurately detect different states and faults as separate clusters is highly dependent on the ability of the dimension reduction technique to extract the inner structure and information from the original high-dimensional data. Therefore, if some states are not distinguished from each other after the first step of dimension reduction, complete fault isolation in clustering is almost impossible. For example, in the visualization of the two-dimensional data, it was observed that data points of Fault 14 were close to those of the normal state or even were indistinguishable in most of the cases. Fault 14 is one of the

Table 3.1: AMI Scores for Cluster Assignments After Dimension Reduction

	Hierarchical	k-Means	k-Medoids	GMM	DBSCAN	OPTICS	CLIQUE
PCA	0.511	0.513	0.455	0.630	0.645	**0.683**	0.373
MDS	0.539	0.510	0.427	0.553	**0.600**	0.501	0.433
ISO	0.655	0.589	0.573	0.440	**0.707**	0.700	0.426
t-SNE	0.766	0.610	0.716	0.642	**0.785**	0.668	0.766
UMAP	0.685	0.595	0.575	0.628	**0.704**	0.673	0.433

Table 3.2: ARI Scores for Cluster Assignments After Dimension Reduction

	Hierarchical	k-Means	k-Medoids	GMM	DBSCAN	OPTICS	CLIQUE
PCA	0.346	0.350	0.292	0.502	0.505	**0.513**	0.165
MDS	0.385	0.413	0.355	0.421	**0.470**	0.287	0.210
ISO	0.513	0.436	0.445	0.209	**0.568**	0.565	0.193
t-SNE	0.635	0.492	0.617	0.545	**0.766**	0.520	0.635
UMAP	0.525	0.450	0.513	0.555	**0.579**	0.555	0.244

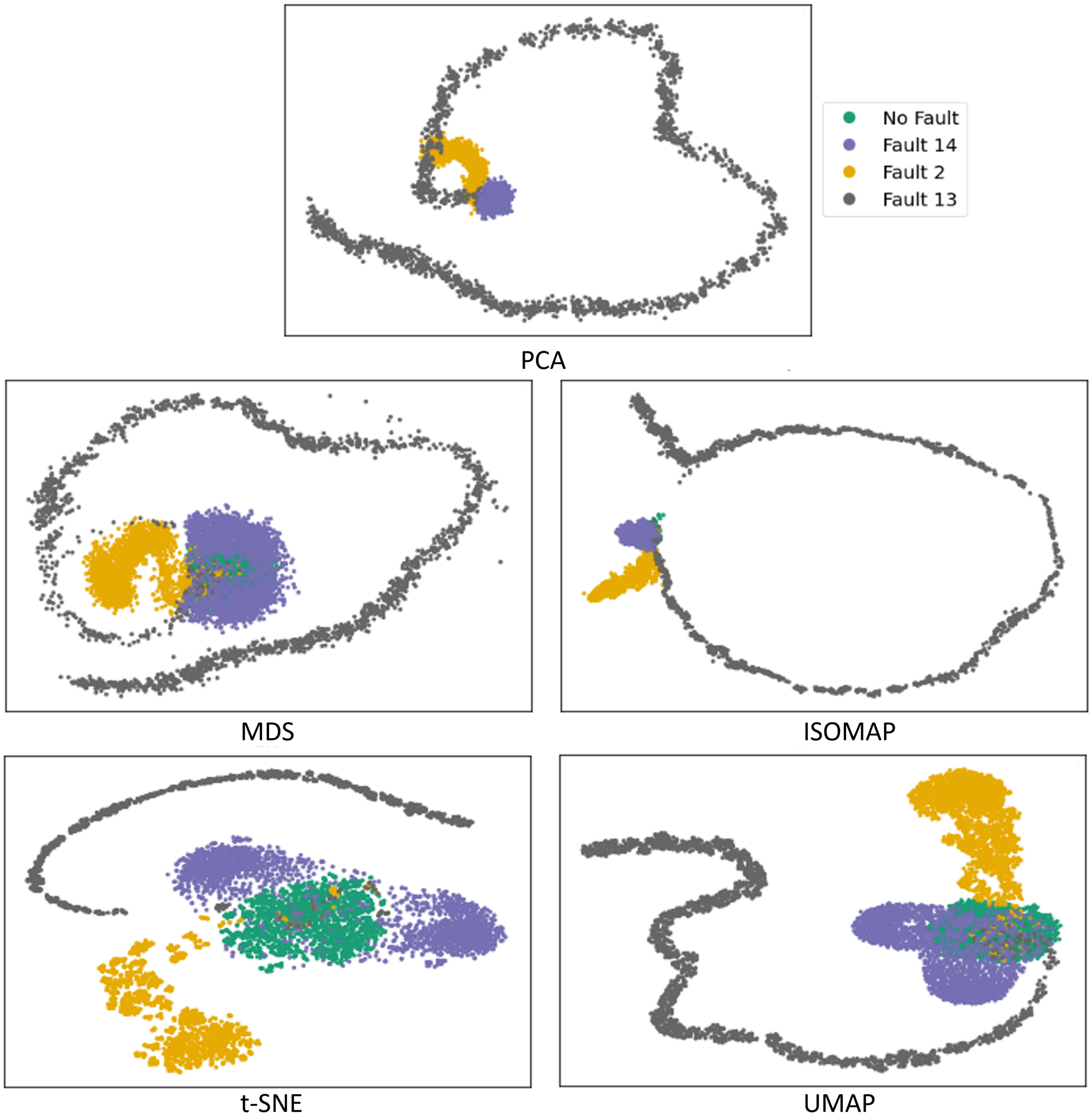
”inconclusive” faults, which are successfully detected only by some methods, and are not detectable easily using others [12]. Two-dimensional projections of this dataset for all DR techniques are presented in subplots of Figure 3.3. In this figure, different colors represent the true labels of the data points for visualization purposes, and no clustering has been performed.

The results of the case study are summarized in Tables 3.1 through 3.3. Table 3.1 presents AMI scores, Table 3.2 presents ARI scores and Table 3.3 displays the V-measure scores. The results demonstrate the obtained scores for each of the clustering methods for each DR technique. The best score for each dimensional reduction technique is highlighted in bold.

A brief overview of the Tables 3.1 through 3.3 shows that the best performing clustering methods for all DR techniques were in the density-based category (the reason is addressed later in this section). To explore in more detail, as presented in Table 3.1, for the correlation-preserving category (e.g., PCA), a density-based clustering had the best performance, and the distribution-based clustering was the method with the second-best performance. For the distance-preserving category (e.g., MDS and Isomap), DBSCAN achieved the highest AMI score for both techniques. The best clustering result for the neighborhood-preserving category was also DBSCAN. For both methods in this category, hierarchical clustering also produced results which

Table 3.3: V-measure Scores for Cluster Assignments After Dimension Reduction

	Hierarchical	k-Means	k-Medoids	GMM	DBSCAN	OPTICS	CLIQUE
PCA	0.512	0.514	0.456	0.630	**0.645**	0.638	0.373
MDS	0.539	0.511	0.427	0.553	**0.600**	0.501	0.433
ISO	0.655	0.589	0.573	0.440	**0.707**	0.700	0.426
t-SNE	0.767	0.610	0.717	0.642	**0.785**	0.668	0.767
UMAP	0.685	0.595	0.575	0.629	**0.705**	0.673	0.434



PCA

MDS

ISOMAP

t-SNE

UMAP

Figure 3.3: Two-dimensional projections of the TEP dataset for all DR techniques

Table 3.4: Performance Rankings of Clustering Categories for Each DR

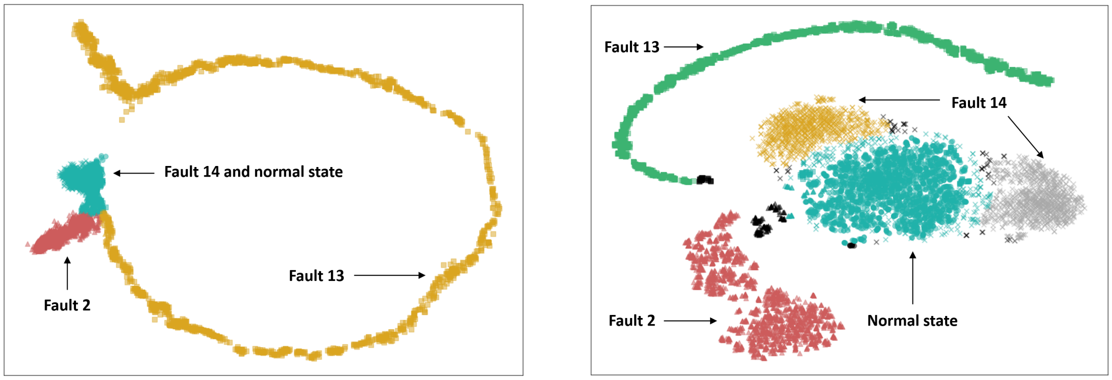
	1 (Best)	2	3	4	5 (Worst)
Correlation-preserving	Density-based	Distribution-based	Distance-based	Centroid-based	Grid-based
Distance-preserving	Density-based	Distance-based	Centroid-based	Distribution-based	Grid-based
Neighborhood-preserving	Density-based	Distance-based	Distribution-based	Centroid-based	Grid-based

were relatively close to the highest scores. These behaviors were also observed in other scores and these results can be confirmed.

To summarize, the performance of each clustering category was ranked for each category of DR and the rankings are presented in Table 3.4, with 1 being the best-performing category and 5 being the worst-performing category.

For all three categories of DR, density-based methods had the best performance and for two out of three categories, distance-based methods ranked second. For all three categories grid-based methods had the worst performance and for two of three categories, centroid-based methods ranked fourth. Grid-based methods and centroid-based methods do not work well with elongated clusters and tend to separate a long cluster into several smaller ones, because of their optimization objectives. In this case, Fault 13 has such a feature as presented in Figure 3.3.

Overall consideration of the results demonstrates that density-based clustering methods lead to higher scores in general, therefore more accurate cluster assignments. Most of the tested DR techniques represent similarity of data points in terms of their closeness, therefore the clusters can be found using a method which searches for areas with high density or clusters them based on the relative distance of data points more easily. An additional observation can be made by comparison of DR techniques, stating that the clustering results obtained after t-SNE demonstrate better performance compared to all other techniques for the dataset at hand.



(a) Isomap Dimension Reduction

(b) t-SNE Dimension Reduction

Figure 3.4: DBSCAN clustering applied on (a) Isomap and (b) t-SNE results

Highest AMI scores were achieved using t-SNE followed by DBSCAN clustering and using Isomap followed by DBSCAN. These results are presented in Figure 3.4. In this figure, different colors represent different clusters found through the mentioned steps, and the true classes of faults are demonstrated using arrows. Three clusters can be seen in Figure 3.4(a). Two faults (2 and 13) were detected successfully and two of the states (Fault 14 and no fault) are almost overlapping and combined into a single cluster in color blue. Figure 3.4(b) shows five different clusters in addition to outliers (in black) found by DBSCAN. All four different states were detected almost perfectly, with the exception of the yellow and gray clusters which are labeled separately but both are Fault 14. It can be seen that Fault 14 is mapped very closely to the normal state with no faults, and Fault 13 has an elongated shape.

3.5 Conclusions

This chapter focused on studying the performance of categories of clustering methods when used in conjunction with categories of DR techniques to enhance the fault detection process. The tests were carried out on a case study from the TEP simulator to find the DR techniques and clustering methods that complement each other. Different combinations of DR techniques followed by clustering methods were applied to the dataset to obtain the highest similarity in cluster assignments to true data labels. The study shows that different categories of DR tend to produce results with specific characteristics leading to their higher compatibility with categories of clustering methods. More specifically, the results show that the density-based and distance-based clustering categories have a promising performance in identifying isolated states and possible faults used along with all DR categories.

4 Performing Multi-objective Optimization alongside Dimension Reduction to Determine Number of Clusters

4.1 Introduction

Constant monitoring of operational variables throughout the systems of manufacturing processes results in massively archived databases. As this historical record of operations is high-dimensional and complex, it presents a valuable opportunity to discover operational patterns and behaviors, and, thus, to detect possible faults and anomalies. In most cases, the adequate analysis of failures or anomalies is only possible after the event, therefore carrying out a post mortem study of historical data becomes critical. As a result, many fault detection methods utilizing high-dimensional historical data have been developed throughout the years and benefited from the use of many data science tools [15, 74].

The historical data collected as such ideally contains information which can be exploited for modeling, prediction and anomaly detection/isolation in various industries. Depending on the size and the complexity of operations, such data can be low-dimensional or high-dimensional, where in the latter, the process of information extraction becomes more challenging. Hence, Big Data analysis has drawn more attention in search of ways to help this knowledge discovery [75]. The approach presented in this study is generic and can be applied to historical datasets from a variety of industries and technologies.

Two tools which are frequently used in information extraction from complex datasets are dimension reduction and clustering methods. Over the years, a variety of approaches have been developed for each of these tools and one needs to be cognizant of the underlying science that makes them appropriate for some datasets and not necessarily others. Often times, a certain amount of insight is required to select the best-performing method, however, this becomes problematic in many real-life cases. Hence, understanding the approach behind each technique can help in guiding the analyst towards the appropriate selection from the available methods. Such lack of clarity led to the use of these methods often in an unsupervised manner [76]. As mentioned previously, DR techniques are one category of highly useful tools which can transform the original high-dimensional dataset to a dataset with lower (thus more manageable) dimensions by preserving specific characteristics of the original data, with the goal to retain the most important features and information. The other aforementioned tool, clustering, serves the purpose of information extraction by finding groups or clusters within the data (classification). These clusters of data points are identified based on the similar characteristics of their members, where each cluster would try to represent a physically meaningful class. The definition of similarity for clustering the points varies among different approaches. These methods have proven to be beneficial in developing data-driven methods for process monitoring and fault detection/diagnosis as compiled by [12] and [10]. After studying the synergy between DR techniques and clustering methods through consideration various combinations and permutations, the matter of obtaining optimal results is a subject that requires further discussion, especially if these methods are used in cases without any prior knowledge about the data.

The most common observation in performing clustering analysis on a dataset is that the odds of obtaining satisfactory results is low, especially if there is not enough prior knowledge about the dataset. This is an outcome of the parametric dependence of the clustering methods on dataset features. In addition, as in most real-life scenarios, since a priori information is not available, performance evaluation can only be done using internal metrics. Most of these validation metrics optimize parameters which are also the objective of the clustering methods [64], so these metrics might not offer a true representation of the data under scrutiny. For example, Silhouette coefficient is more compatible with k-means and Density-Based Clustering Validation (DBCV) index is more compatible with HDBSCAN. Therefore, evaluating internal validation criteria is a matter which has been examined through various studies to measure the ability of these metrics in representing the success of clustering results [77].

Multiple studies have been performed in the last few decades addressing the performance of internal validation criteria, each focusing on different information to report, such as the ability to identify the correct number of clusters, assessing the correct grouping of data points and selecting the best parameter settings for each algorithm [77]. One of the first studies was carried out by [78], and 108 small, artificial datasets were used to identify the correct number of clusters for hierarchical clustering. They reported that Calinski-Harabasz (CH) index was the best-performing one. In a more recent study by [79], four validation criteria, Silhouette, Davies-Bouldin, CH, and DBCV Index were used on 27 datasets to identify the highest scoring solution and the best parameters for each of the six clustering algorithms; k-means, DBSCAN, Ward, Expectation-Minimization (EM), mean-shift and spectral clustering. Their results indicate higher compatibility of specific clustering methods with some internal metrics compared to others.

When faced with multiple indices in assessing performance, a reasonable approach would be to consider an ensemble of these indices. Then, naturally, the matter of partial weights and metrics combinations arise. Another approach would then lie in exploring a framework in which more than one index is considered at a time to discover the patterns underlying the data, hence the best possible clustering solution. As a result, it seems appropriate to formulate the problem as a multi-objective optimization problem. Multi-objective optimization, or Pareto optimization, is an optimization problem which has, as the name suggests, more than one objective function to optimize. Pareto optimal or non-dominated solutions are the solutions in which none of the objective functions can be improved without degrading another one, thus producing the set of most acceptable solutions. Realistically, this also means that there exists no solution that has every objective simultaneously optimized (maximized or minimized). Therefore, in the end, a set of solutions are deemed equally acceptable and presented to the decision maker [80]. There are three major approaches to multi-objective optimization problems: (i) using a weighted sum to combine a set of objectives into a single objective, (ii) ordering the objectives based on their importance, or (iii) optimizing the multiple objectives simultaneously with a Pareto based approach [81]. The latter provides a set of solutions from which the decision maker will select.

Multi-objective clustering (MOC) has become of growing interest in the recent years, especially in the field of genetic algorithms [82]. Keeping in mind that, realistically, optimizing no single objective can completely capture the clusters present in a dataset, MOC performs a search over the search space of all parameters of the algorithm while simultaneously optimizing a number of validity indices. An algorithm named MOCK (Multi-Objective Clustering with automatic determination of the number of clusters) presented by [83], offers a multi-objective evolutionary algorithm to detect the correct number of clusters in a genetic encoding, and its objective functions are *overall deviation* of partitioning (which is the overall summed distances between data points and their cluster centers) and connectivity (which measures the degree that neighboring data points have been placed in the same cluster). In another work performed by [84], NSGA-II algorithm has been used on numeric image datasets to simultaneously optimize the Xie-Beni (XB) index (which is a ratio of global to local variation) and J_m (which calculates the global cluster variance), and the results are compared to two other scenarios, each optimizing only one of these scores using other clustering methods. The final results are evaluated using another index which is not included before, and suggests that the multi-objective clustering outperforms the other methods.

Considering the previous work and aforementioned advantages of MOC, this chapter considers a clustering strategy using Pareto optimization to simultaneously optimize a number of different metrics to reveal the behavior of the data. Hence, the clustering result will not depend on only one metric (and one type of cluster feature), but a more holistic view is taken into account to assign the memberships. In more detail, the clustering would be treated as an unsupervised step of data analysis, with none of the parameters selected. The clustering is performed multiple times while varying the parameters over their ranges, and multiple internal metrics are calculated for each run. A multi-objective optimization is then performed, with these metrics as objectives, to find the best solutions and the final number of present clusters. To get a better insight, the previous step is then combined with DR techniques to further assist the information extraction. Depending on how the mapping is carried out from high dimensions to low dimensions, different clusters can be created or lost. Therefore, considering representations of the dataset at hand in various versions in lower dimensions would help in identifying the patterns/clusters which are preserved during the mapping, i.e., the most persistent features of the data. Finally, MOC is performed multiple times on different projections of the data to assess the results of the clustering step.

In comparison to previous studies focusing on multi-objective clustering, the novel features of the current study are the consideration of a density-based validity metric as well as cohesion-based and separation-based indices simultaneously, as well as the utilization of DR techniques and application of multi-objective clustering in tandem to discover the most important features of a dataset. In the next section, problem formulation, details of the datasets for the case studies and the steps of the proposed approach are presented. Results and discussion of the numerical experiments on the mentioned datasets are followed, and in the end the conclusions are offered.

4.2 Methodology

In this section, the general mathematical formulation of the problem is presented, followed by the utilized methods, then the datasets are introduced. The section concludes with the specific steps of the proposed method.

4.2.1 Problem formulation

The schematic representation of the proposed methodology is depicted in Figure 4.1. To formulate the problem, let us assume a $n \times m$ data matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$, where n denotes the number of samples and m is the number of variables. During the first step, the aim of dimension reduction is to map the matrix $\mathbf{X}$ onto $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T$ where the set of vectors is of a reduced dimension p ($p \leqslant m$) by preserving as much of the intrinsic structure of the data set as possible:

$$\mathbf{Y} = F(\mathbf{X}) \tag{43}$$

In this expression, the function F can take different forms depending on the selected dimension reduction technique.

Next, let us assume there are n_L distinguishable categories/clusters in the data set, and let them represent different operational states, faults, chemical samples, etc. dictated by the features expressed by the data set. Clustering aims to categorize (classify) the data samples into these n_L separate subsets in an unsupervised manner, aiming to group similar instances together, while different instances should belong to different categories (classes). In the present work, clustering is performed on the $\mathbf{Y}$ matrix of reduced dimension, however, it must be noted that it can be performed on the original data set as well, at least for purposes of comparison. This step is represented as,

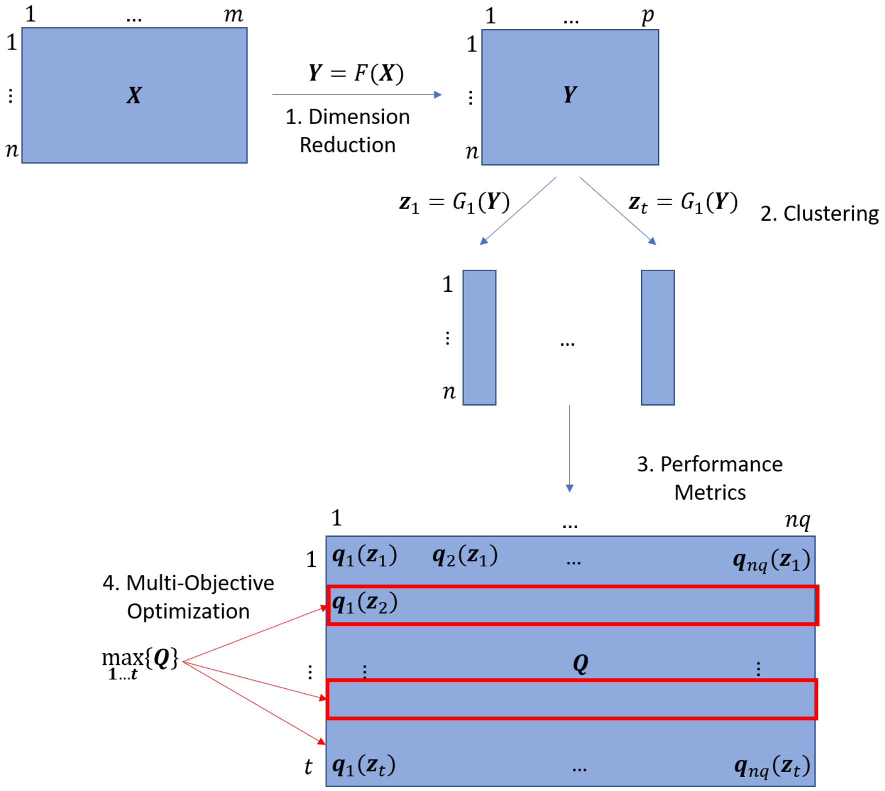


Figure 4.1: The schematic representation of the proposed methodology with its four main steps: dimension reduction, t types of clustering, the characterization of the clustering solutions by nq metrics and the selection of the best solutions

$$\mathbf{z} = G(\mathbf{Y}) \tag{44}$$

Again, the function G is defined based on the selected method. Here, $\mathbf{z}$ is a $n \times 1$ vector, where the elements correspond to the subsets $\mathbf{C} = \{C_1, ..., C_i, ..., C_k\}$ with the number of clusters k. In addition, it is assumed that there exists true set of labels defined by a $n \times 1$ vector $\mathbf{u}$, where the elements are of the subsets of true labels $\mathbf{L} = \{L_1, ..., L_j, ..., L_{n_L}\}$ with the number of labels represented by n_L.

Recognizing that clustering is an unsupervised machine learning technique, the determination of the number of clusters is often based on the expert knowledge of the application domain and in several cases may be required as an input to the clustering algorithm. The aim of the present chapter is to determine the optimal number of clusters, k, in an unsupervised manner. Therefore, during step two, t different possible clustering solutions are created, resulting in a total of t vector $\mathbf{z}$'s. Subsequently, nq internal metrics are calculated for these solutions in the third step and stored in a $t \times nq$ matrix $\mathbf{Q} = [Q_1(\mathbf{z}), Q_2(\mathbf{z}) \dots, Q_t(\mathbf{z})]^T$. This is followed by the determination of the optimal solutions from the t solutions and further analysis reveals the optimal number of subsets, where the number of found clusters and the number of true (distinguishable) subsets in the data set are equal, i.e., $k = n_L$. Now the general formulation of the multi-objective optimization problem in the fourth step with nq objective functions can be presented as,

$$\max_t \{Q = [q_1(\mathbf{z}), q_2(\mathbf{z}) \dots, q_{nq}(\mathbf{z})]\} \tag{45}$$

Due to the often conflicting objective functions, the solution of a multi-objective optimization problem is typically not a single point but a group of points called the Pareto-optimal or non-dominated set of solutions. A dominated solution is defined as the solution which is strictly better than the rest of the solutions in at least one criterion and is no worse than the rest in all objectives. In other words, non-dominated solution cannot improve in any of the objectives without degrading at least another one, and are accordingly selected as a member of the final solution set. This set of non-dominated solutions is called Pareto-optimal [85].

In the present work, the goal is to analyze the Pareto optimal solutions of the t solutions that are found through clustering. The form of the applied dimension reduction functions, F, the approaches for clustering, therefore, the form of the function G and the objective functions, or in other words, the internal evaluation metrics (indices) of clustering performance, are provided in Chapter 2, as well as external metrics for result assessment of the proposed approach.

4.2.2 Utilized methods

Dimension reduction techniques

The main use of DR techniques is to remove redundant information (data) and represent meaningful features of the raw process data in fewer, often, two or three dimensions. Some benefits of applying DR techniques are reducing the computational complexity and avoiding the curse of dimensionality. While having a lower number of (transformed) variables facilitates data analysis, these techniques can also be used for visualization purposes in 2D or 3D. For the case at hand, all DR techniques have been used to transform the original m-dimensional dataset into a two-dimensional manifold. All DR techniques introduced in Chapter 2.2, 5 techniques in total, are used in this part of study, with focus on their definitive attributes which classifies them into three different categories of correlation-preserving, distance-preserving and neighborhood-preserving.

Clustering methods

All clustering methods introduced in Chapter 2.3 in any of the four different categories of clustering methods of connectivity-based, centroid-based, distribution-based and density-based are utilized in this part of study, with a total of 7 methods.

It is important to note that the performance assessment of clustering methods depends on an important piece of information about the data; whether the true assignments of data points (i.e., true labels) are known or not. Cases with known true labels are usually used for training and classification (supervised) purposes and metrics are called external. There are a number of important metrics which are calculated using the true labels and clustering assignments, as introduced in Chapter 2.3.6. The ones considered in this study are Adjusted Mutual Information (AMI) and V-measure. These indices provide a more accurate assessment of how the clustering was performed. For the purposes of this study, external metrics are only introduced and later utilized to demonstrate the performance of the approach, and are not objective functions for the optimization process. However, for the reasons mentioned previously, the performance assessment of the clustering assignments is carried out with the help of internal features of the found clusters. Internal metrics such as Silhouette coefficient, Davies-Bouldin index and DBCV index are the objective functions of optimization to find the expected number of clusters and are the calculated metrics for each solution.

Multi-Objective (MO) optimization

While Single-Objective (SO) optimization is easy to implement, in most cases the problem contains more than one influential objective to make a decision. As opposed to SO optimizations in which only a single function is aimed to be optimized, MO optimizations can be challenging. These objectives often can be conflicting, hence, resulting in a set of solutions known as Pareto-optimal solutions. Any two solutions from this set present a trade-off between the objectives, therefore, the selection of the final solution requires the decision maker's knowledge and further investigations. Optimization in the context of this research can be defined as the process of finding the best solution(s) possible from the available ones [86]. Therefore, based on some criteria, multiple objective functions are defined. Then, a search algorithm is utilized to maximize the objective functions. Realistically, there is no unique solution which optimizes all objective functions at the same time. Hence, a set of solutions are presented in the end with an acceptable level of trade-off [80]. This set of solutions contains two groups of dominant and non-dominant solutions. A solution is considered dominant if:

1. Is no worse than other solutions in all objectives

   and

2. Is strictly better than other solutions in at least one objective

Therefore, MO optimization is performed using the concept of dominance. If a number of solutions are non-dominant, meaning one of the conditions above is violated, one cannot decide on the supremacy of the solutions. Each solution may be optimizing one of the objective functions and performing poorly in the other, and since all objective functions are important, all such solutions are deemed acceptable. Therefore, the set of Pareto-optimal solutions are the non-dominant ones [86]. In this study, objective functions are internal clustering metrics and the solution set are the clustering labels found for each clustering method. Therefore, to perform MO optimization. following algorithm has been used:

- Start enumerating through all members of the decision space
- Compare the solution with all the ones after it for domination:
  - If later solutions dominate the current one, keep them
  - If not, remove any solution that is not better in at least one objective than the current one

The algorithm leads to a set of non-dominant solutions which can be equally accepted. It is important to note that the solutions which are optimizing the objective functions might not be the guaranteed optimal solutions, yet they are the best solutions which could be selected [86].

4.2.3 Datasets

Three different datasets with distinct features were tested using the approach presented later in section 4.2.4.

- **Wine dataset**

  The first dataset is the well-known and broadly used "Wine dataset", containing the results of the chemical analysis of wines grown in a specific area of Italy, available on the UCI repository [87].

The dataset contains three different wine types represented by overall 178 samples, with 13 variables recorded for each sample. Based on the dataset, it is desirable to detect three individual groups after performing clustering, i.e., when using clustering methods that do not perform outlier detection, the *expected number of clusters* is three.

- **Synthetic dataset**

  The idea to explore in the second dataset is to evaluate the performance of this approach on 2D datasets. The dataset is synthetic and created using *make_blobs* function of the scikit-learn package [88]. It contains 5 clusters, each with a 0.6 standard deviation and is in 2D as mentioned. There would be no dimension reduction in this case, but the DR techniques are used still to produce different projections of the data. In this case, using the clustering methods that do not detect outliers, the *expected number of clusters* would be five.

- **Fault detection in the Tennessee Eastman Process**

  Third dataset is generated using the revised Tennessee Eastman Process (TEP) simulator introduced in Chapter 3.2, and is the same dataset utilized in the previous chapter. As an overview, TEP is a benchmark chemical plant consisting of five major units (a reactor, a product condenser, a recycle compressor, a vapor-liquid separator, and a product stripper) and eight chemical components. For this dataset, three different faults of varying types were selected; Fault 2 is a step change in a component composition; Fault 13 is a slow drift in the reaction kinetics, and Fault 14 represents the sticking of the reactor cooling water valve. Each of these faults were separately active for a constant 20 minutes, and then turned off right before the next period started. The dataset also contains 20 minutes of normal operation (without any faults).

  As demonstrated in Chapter 3.4 in the visualization of the 2D data, it was observed that data points of Fault 14 were close to those of the normal state or even were indistinguishable in most of the cases, which makes it difficult for the clustering methods to detect two different states. During the mapping from high dimensions to low dimensions, these two states are considered to have similar features, hence, they are overlapping in the lower dimensions. Therefore, if it can be deemed acceptable for the clustering methods to consider Fault 14 and normal state as 1 cluster because of the DR performance, then the *expected number of clusters* can be considered 3. Also, as demonstrated, the average number of points per cluster in this dataset is higher compared to previous datasets and is much higher than the possible outliers, because of the simulation runs. In other words, there are not many outliers present in the data, and the *expected number of clusters* would be 3 for any method.

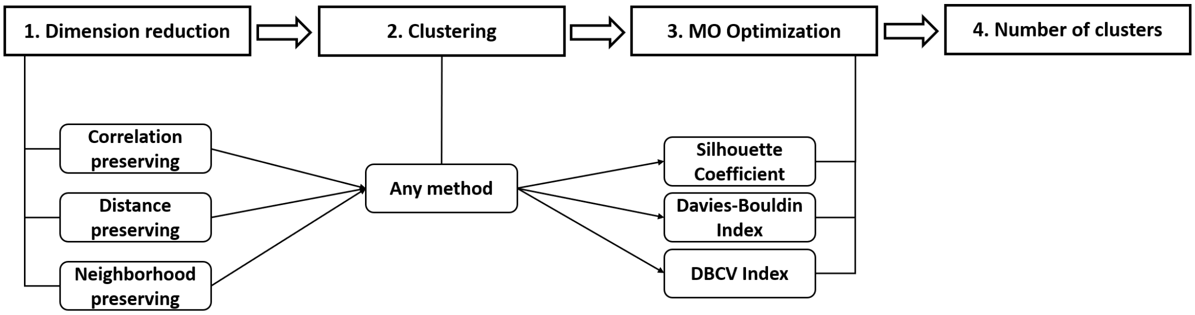


Figure 4.2: A step-wise summary of the proposed method to detect the number of clusters

One important feature of this dataset is the presence of an elongated cluster belonging to Fault 13, which is a slow drift in the reaction kinetics. As mentioned previously, methods such as the ones in centroid-based category do not perform well when encountered with elongated clusters. Hence, methods such as DBSCAN or HDBSCAN, would be expected to perform better for datasets with elongated clusters, such as this one.

4.2.4 Proposed method

To start with, after normalization of the data, multiple lower-dimensional forms of the dataset are created using different DR techniques, while also keeping the original high-dimensional version. Then, for each clustering method discussed above, a clustering analysis is performed over the search space of its hyper-parameters. This is followed by Pareto optimization to choose the best set of solutions while optimizing the internal metrics, hence selecting the best number of clusters without any prior knowledge. Due to the varying nature (and goals) of clustering methods, it is not meaningful to compare the final results, since methods have different capabilities in terms of outlier and cluster shape detection; some methods detect outliers while others are unable to detect them, and some methods cannot work with elongated clusters while others are more compatible with different shapes of clusters. Therefore, one universal element which can be compared over all clustering results is the final number of clusters found. Hence, that will be the focus of the comparative study. Figure 4.2 presents a roadmap of the steps included in the current approach.

The process of transforming the dimensions is not at study here. It is assumed the dimension reduction with suitable parameters is feasible and the parameter selection for DR (if needed) is performed based on available studies, or other heuristics. All the dimension reduction techniques have been used to obtain 2D datasets. All five mentioned DR techniques have been applied paired with each clustering, so for each dataset there are 6 versions available in total. More DR techniques can be used, and this number has been selected as a minimum. Also, all four categories of previously mentioned clustering (seven methods in total) were utilized for a more thorough study.

As mentioned, in order to emphasize the unsupervised nature of this study, only internal metrics are considered for optimization and calculation of any external metrics is solely for the purpose of result comparison. These internal metrics are Silhouette coefficient, DB and DBCV indices. The first two are to be maximized but for the latter metric, lower values indicate better performance. Moreover, the DB index, in the context of this study, is in the order of 1000, while other indices are between 0-1. Therefore, the reciprocal of the DB index is maximized in the optimization process to uniformly maximize all the metrics in the same range.

After Pareto optimization, the number of presented solutions varies among different clustering methods and different dimension-reduced datasets. In some cases, the number of Pareto solutions is very high and sometimes only a handful of solutions are acceptable. Hence, in order to balance the number of presented solutions in each case, and remove the solutions with the poorest performance, a filter is placed after the optimization step. This filter removes the solutions which have any internal score with an absolute value lower than 0.01. Since the closeness of the calculated scores to zero is an indicator of their poor performance, this value has been selected as a heuristic threshold to eliminate any possible substandard solutions.

4.3 Results

In this section, the results of the study are presented in individual graphs. For each dataset, the results are grouped based on the category of clustering methods (distance-based, centroid-based, distribution-based and density-based), resulting in four subplots. Each plot demonstrates the fraction of Pareto solutions representing different number of clusters. The plots contain the performance of each DR technique and an overall performance, considering all the Pareto front solutions found for the dataset. In the end of each part, an example of the actual clustering with the calculated AMI and V-measure is presented. Final part of this chapter is dedicated to key findings and discussion of the results. Details of results containing all dominant and non-dominant solutions for every dataset and each clustering method are available in Appendix C.

4.3.1 Wine dataset

As an example of the optimization outcome, a three dimensional space is presented in Figure 4.3, which summarizes the solution set for the PCA-reduced wine dataset, with metrics calculated using DBSCAN clustering. The figure shows all the solutions obtained for optimizing the three metrics, i.e., Silhouette coefficient, Davies-Bouldin index and DBCV index. The solid dots are the dominated solutions and the points marked with x are the non-dominated solutions (Pareto front). The threshold regions to remove the

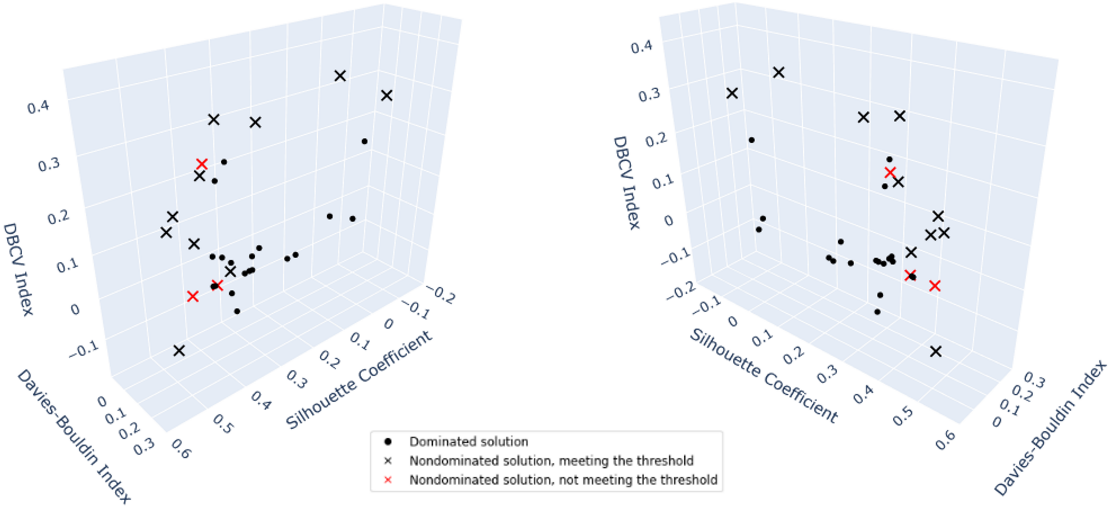


Figure 4.3: A three dimensional representation of the optimization space.

solutions with poor performance resulted in removing the points which are marked in red. As demonstrated, in this case, three non-dominated solutions were removed, and the rest of the Pareto front is considered in the cluster number detection to calculate the fraction of solutions suggesting each number of clusters.

Different high-dimensional and low-dimensional versions of the dataset suggest different number of clusters as their most repeated solution, while it should be kept in mind that each version presented a different number of Pareto front solutions. For example, as shown in Figure 4.4, for the distance-based category, all solutions found after ISOMAP dimension reduction suggested that there are two clusters present and t-SNE results indicated that there are equal possibilities for the presence of 3 clusters and 4 clusters. Therefore, considering all the solutions overall leaves out the effects of each individual DR technique, and helps to examine the features of the dataset which persisted throughout all the high-dimensional and low-dimensional versions, i.e., the number of clusters.

Figure 4.4 indicates that, considering distance-based clustering methods, which includes agglomerative hierarchical clustering, 0.53 of the solutions suggest there are 3 clusters in the dataset. For the centroid-based clustering, i.e., considering both k-means and k-medoids, this number is 0.44 and for the distribution-based category including GMM, this fraction is 0.46. Density-based methods, DBSCAN, OPTICS and HDBSCAN considered all together, suggest 4 clusters in 0.60 of their solutions. It can be seen that first three categories of clustering methods detected 3 clusters, *expected number of clusters*, and the last category detected 4 clusters. It should be noted that the methods in the final category are able to isolate outliers, hence ending up with

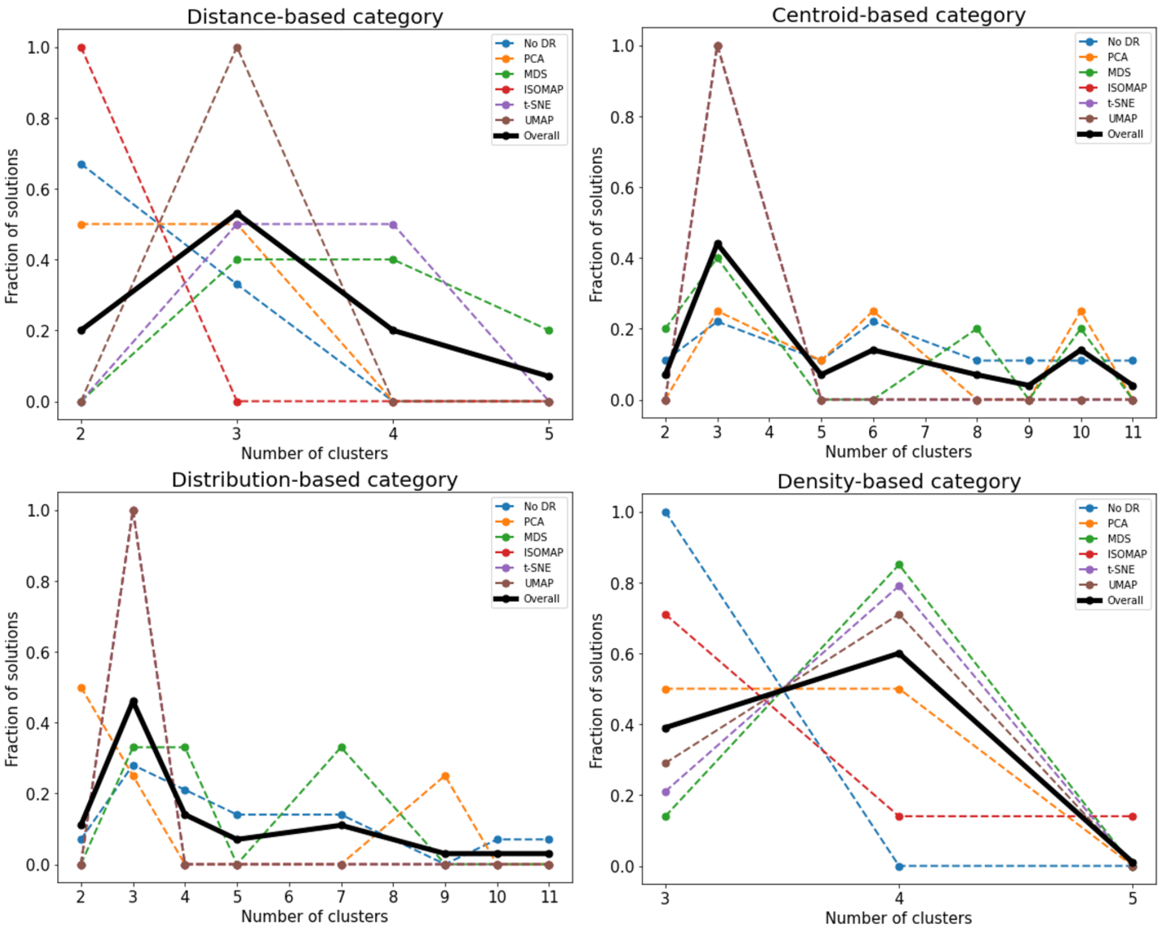


Figure 4.4: Fraction of solutions representing found number of clusters in the wine dataset for each DR technique.

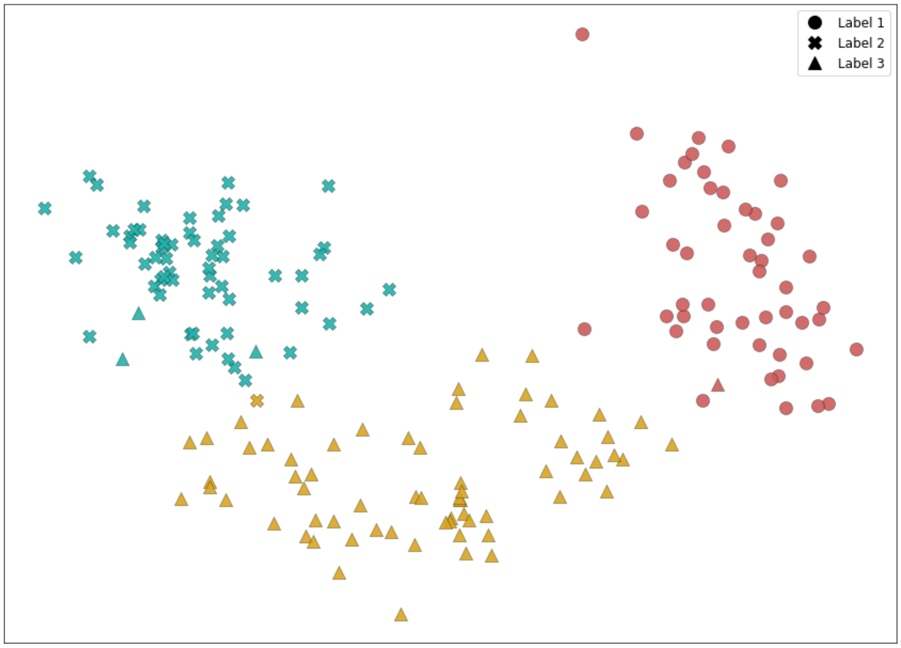


Figure 4.5: GMM clustering of the wine dataset with ISOMAP.

expected number of clusters+1. Although no outliers were reported explicitly in this particular dataset, some points were not assigned to any of the clusters, suggesting this possibility.

The selection of a solution(s) from the Pareto set is called post-Pareto optimality analysis [89], which is not pursued in this work, but to observe an example of the clustering performance itself, one solution is displayed in Figure 4.5. The dataset is converted to two dimensions using Isomap, and clustering is performed using GMM. Colors represent clusters found by the method, and the true labels are marked with different shapes. The AMI obtained by this solution is 0.833 and V-measure is 0.835, indicating a satisfactory outcome.

4.3.2 Synthetic dataset

For this dataset, Figure 4.6 shows how each clustering method performed in the determination of the number of clusters. The distance-based category suggests 5 clusters with 0.50 of the solutions, the centroid based suggests 5 clusters with 0.46 of the solutions, and distribution-based category indicates 5 clusters with 0.38 of the solutions. In all these cases, the cluster number with the highest fraction of the solutions does not have the majority (more than half), but it is the most-repeated outcome compared to others. On the other hand, the density-based category suggests 6 clusters with 0.76 of the solutions. The first three categories correctly detect 5 clusters in the dataset, same as the *expected number of clusters*. For the last category (three methods), the proposed approach suggests that there are 6 clusters in total, including the cluster of outliers. Again, the final results are consistent with the *expected number of clusters+1*. In the first three plots there are no solutions with more than 5 clusters, meaning that no non-dominated solutions detected more than 5 clusters.

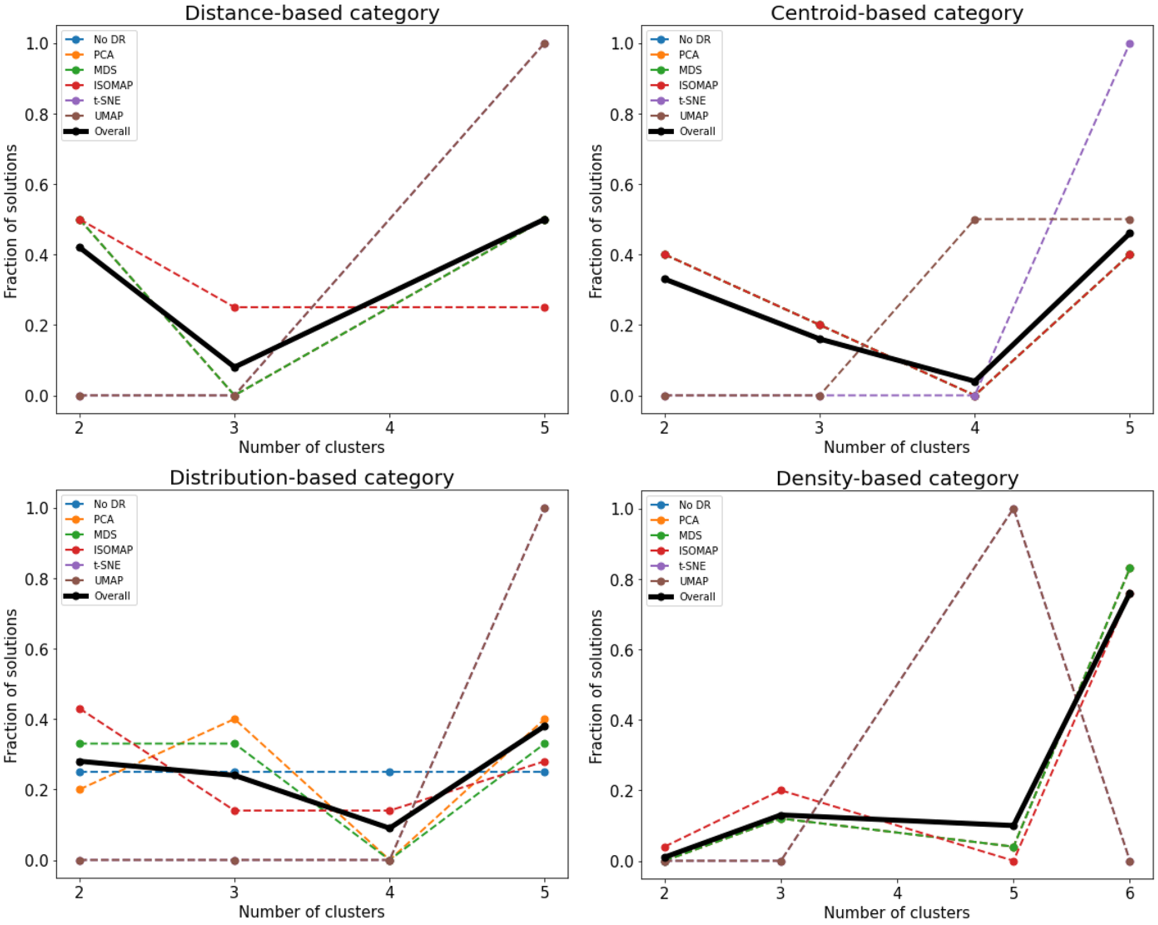


Figure 4.6: Fraction of solutions representing found number of clusters in the synthetic dataset for each DR technique.

An example of clustering performance for this dataset is demonstrated in Figure 4.7. The DR technique is MDS and clustering is performed using DBSCAN. The five obtained clusters are in different colors, with the outliers in gray. The true labels are marked with different shapes, and the AMI for this clustering result is 0.835 and V-measure is 0.835.

4.3.3 TEP dataset

For the final dataset, it can be seen that the distance-based category suggests the presence of 3 clusters or 9 clusters, each with a 0.16 probability, the centroid-based suggests 8 clusters with 0.16 probability, the distribution-based category suggests 8 clusters with 0.16 and 3 clusters with 0.15 probability. In the

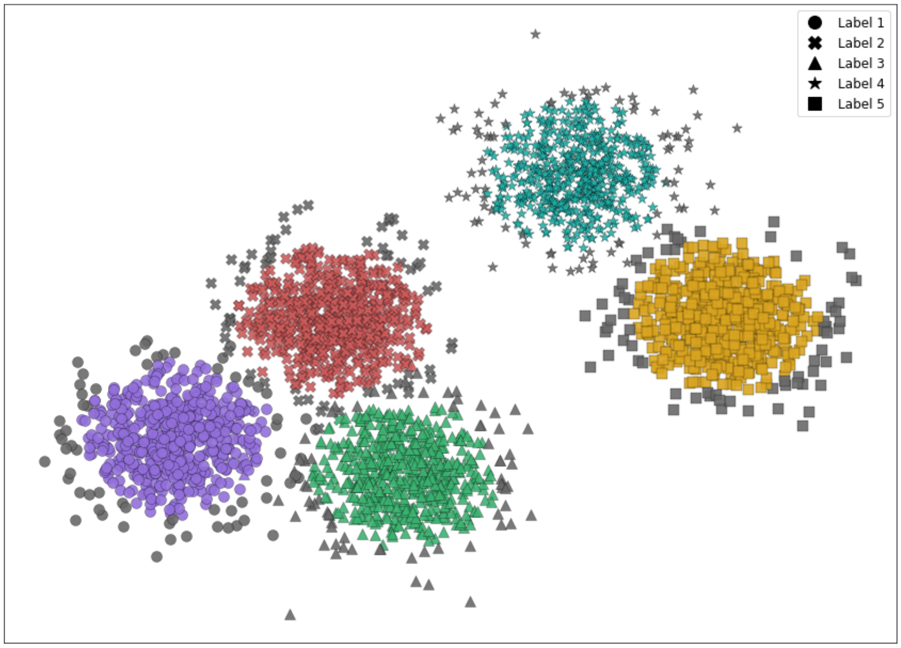


Figure 4.7: DBSCAN clustering of the synthetic dataset with MDS.

last category, the density-based, the final number is presented with a higher confidence compared to other categories, which is 3 clusters with 0.37 of all solutions.

Based on what was noted for the *expected number of clusters* for this dataset, the density-based category performed well. Their final results indicate 3 clusters which is equal to the *expected number of clusters*. Another category which suggested the presence of 3 clusters in the dataset, but with a lower confidence, was the distance-based category. This is in confirmation with the results of [32] demonstrating this category had the second-best performance in state isolation. From one of the solutions, the clusters found using hierarchical clustering on the t-SNE reduced data are presented in Figure 4.9, with an AMI of 0.766 and V-measure of 0.767. As it can be seen in this figure, the normal state and Fault 14 are combined in one cluster in yellow.

4.3.4 Discussion

Overall, this method demonstrates an outstanding performance in the determination of the number of clusters present in a dataset. Key observations and further discussion of the results are listed below:

- Using any method which is not from the density-based category, the *expected number of clusters* is obtained. In cases where density-based category detect one cluster more than the other categories, the final number can be interpreted as *expected number of clusters+1*.

- In cases where different categories present inconsistent results, there is a possibility that an elongated

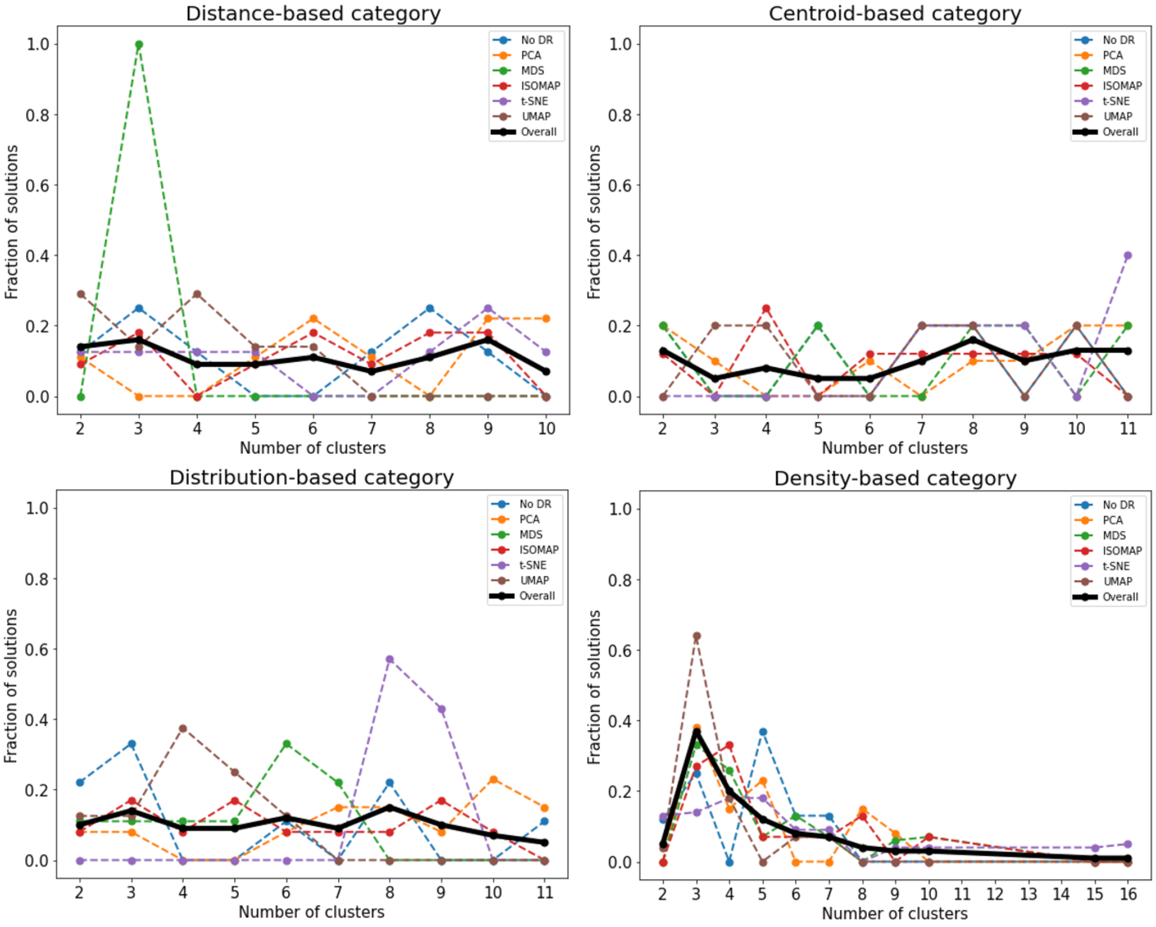


Figure 4.8: Fraction of solutions representing found number of clusters in the TEP dataset for each DR technique.

cluster is present. In this case, further examination of the original dataset and its lower-dimensional versions is needed to choose the final number of clusters.

- Although, with all the attempt to emphasize carrying out all the steps in an unsupervised manner, in the end, the selection of the final number depends on a level of familiarity with the dataset and/or the method as mentioned. A basic knowledge is required to decide whether the final number is with or without the outliers depending on the method, and whether any number of clusters have overlapped during the DR or not. It should be noted that the required level of information can be easily obtained by visualization of the dimensionally-reduced versions of the dataset, since these features can be pointed out in the comparison of the different versions.

- Despite the advantages of overall consideration of solutions in finding the number of clusters in a dataset, it can be interesting and beneficial to look at the individual performances of DR techniques in detecting the clusters. As seen in Figure 4.4, for all categories of clusterings, MDS, t-SNE and UMAP

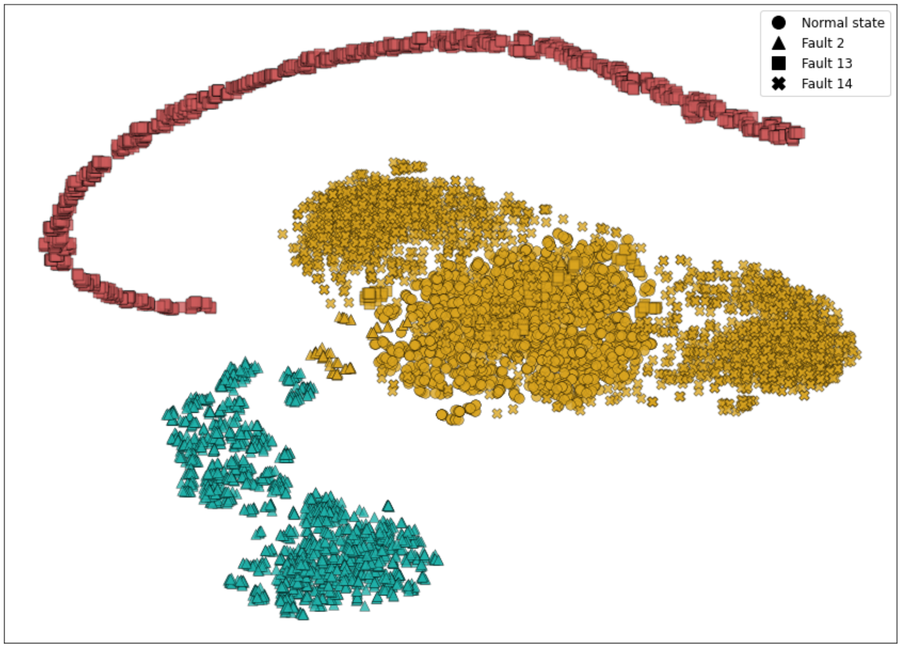


Figure 4.9: Hierarchical clustering of the TEP dataset with t-SNE.

were the DR techniques which suggested the *expected number of clusters*, either with the highest probability or one of highest probabilities. For the synthetic dataset in Figure 4.6, it is demonstrated that t-SNE and UMAP lead to the same results as the overall performance in three out of four clustering categories, with the exception of density-based category. Performance of PCA and MDS is consistent with the overall performance in another three out of four clustering categories, with the exception being distribution-based methods. For the case with elongated cluster, by looking at the Density-based category in Figure 4.8, PCA, MDS and UMAP suggest the same number of clusters as the overall performance. In conclusion, UMAP can deliver good results in almost every case. In the quantitative survey on dimension reduction techniques by [20], UMAP was also suggested as one of the techniques yielding the best quality of projection based on their considered quality metrics.

- Looking at the no DR solutions obtained for each clustering category, there is no specific pattern of behavior in the results of the wine dataset and TEP datase, and the clustering of the original dataset was only able to detect the number of clusters in at most two categories, still not with a high probability compared to other numbers. However, the clustering of the original dataset for the synthetic data lead to the same result of the overall performance. This can be explained by the dimensionality of the data at hand. Synthetic dataset was two-dimensional while the other dataset were high-dimensional, and in the latter, clustering methods utilizing traditional distance measure, such as Euclidean, may be ineffective in detecting the clusters, since such distance measures may be dominated by noise in many dimensions [3].

4.4 Conclusions

In exploring and extracting information from high-dimensional datasets, clustering plays an important role. Most of the clustering methods need some information about internal features of the dataset, if not the actual number of clusters. In order to utilize clustering as an unsupervised data analysis tool, a method has been proposed in this study to simultaneously optimize a number of internal cluster validation metrics to detect the number of clusters present in the data. This process is also carried out on dimensionally reduced versions of the dataset, since it is assumed that the most important features of the data will preserve after dimension reductions. This approach was carried out on three datasets with different features. In all three cases the approach was able to correctly detect the expected number of clusters, and is important to mention that the results were achievable irrespective of the selection of clustering method. Overall, this method can be a useful step to detect the number of clusters in a dataset as it requires none or basic preliminary examinations of features of the data.

5 Conclusions and Future Work

Plant-wide process monitoring provides opportunities for developing state-of-the-art methods and improving current models for fault detection and diagnosis by exploiting the collected data from the chemical plants. Data mining tools are the basis of these data-driven approaches using historical process data. Each of data mining tools have distinct inherent characteristics, and therefore, have different abilities in detecting different patterns in datasets. Although some of these tools, such as clustering methods and dimension reduction techniques have an excellent performance in information extraction individually, combinations of them can increase the performance for further investigations. Therefore, it is necessary to study combinations of clustering and dimension reduction techniques to assist fault detection processes.

In this thesis, a wide variety of dimension reduction techniques and clustering methods are selected to study. Different pairings are tested in conjunction to find the DR techniques and clustering methods which are more compatible with each other for state isolation purposes. Tests were carried out on a dataset from Tennessee Eastman Process simulator and the results indicated a higher performance in state isolation could be achieved by judiciously pairing dimension reduction techniques and clustering methods. Overall, t-SNE was the DR technique with the highest compatibility with all clustering methods, and density-based and distance-based clustering categories had the best performance in identifying faults when combined with all DR categories.

As another part of the research, the ability of dimension reduction techniques in preserving and extracting the underlying structure of a dataset has been studied. Each DR technique has different abilities in extracting these features, therefore, a comparison of results of different DR techniques along with the original dataset can help identify the features that are common in all of them, hence detecting the most important characteristics of the data. This comparison is performed using clustering methods in order to identify the number of states that are detectable without any prior knowledge. In this process, multi-objective optimization has been used to assist with the clusterings. Several internal metrics are simultaneously optimized to find the clusters in an unsupervised manner. This approach performed successfully on three datasets with distinct features, and demonstrated an excellent capability in finding the expected cluster number irrespective of the clustering method and the type of dataset.

Considering the findings of this research, a number of ways can be noted for future studies and investigations. Although the results of this research contributes to improvements in fault detection and diagnosis, it is important to propose potential future directions to build on this basis and lead to further enhancements. The first is to study the solutions and clustering labels suggested by the Pareto front in cluster number detection process. More specifically, a study on the quality of the individual solutions and how to select the most appropriate one is valuable. Inspection of all final solutions in order to select the most appropriate one can be exhausting and not necessarily simple. Therefore, utilizing data mining and machine learning methods for the purpose of searching for patterns is beneficial. Also, comparison of different clustering solutions found using different methods may be helpful in this process. In case of discovering similar clusters from different clustering methods, the presence of a particular cluster can be confirmed. For example, if a number of clustering methods have proposed the same labeling for a dataset, cluster assignments can be accepted with a higher confidence.

Another direction for future research can involve studying additional dimension reduction techniques. Although it has been tried to utilize at least one DR technique from each of the three categories, using other DR techniques can be insightful and helpful to detect the number of clusters with a higher certainty. Moreover, studying different combinations of DRs can lead to finding the optimum number of techniques and/or finding the techniques which can sufficiently provide the required information without performing excessive studies. The same direction can be pursued with studying additional internal metrics.

References

[1] M. S. Mahdavinejad, M. Rezvan, M. Barekatain, P. Adibi, P. Barnaghi, and A. P. Sheth, “Machine learning for internet of things data analysis: A survey,” *Digital Communications and Networks*, vol. 4, no. 3, pp. 161–175, 2018.

[2] J. H. Lee, J. Shin, and M. J. Realff, “Machine learning: Overview of the recent progresses and implications for the process systems engineering field,” *Computers & Chemical Engineering*, vol. 114, pp. 111–121, 2018.

[3] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques.* Elsevier, 2011.

[4] M. K. Gupta and P. Chandra, “A comprehensive survey of data mining,” *International Journal of Information Technology*, pp. 1–15, 2020.

[5] S. Patel and H. Patel, “Survey of data mining techniques used in healthcare domain,” *International Journal of Information*, vol. 6, no. 1/2, pp. 53–60, 2016.

[6] K. Gibert, J. Izquierdo, M. Sànchez-Marrè, S. H. Hamilton, I. Rodríguez-Roda, and G. Holmes, “Which method to use? an assessment of data mining methods in environmental data science,” *Environmental modelling & software*, vol. 110, pp. 3–27, 2018.

[7] D. Namiot and M. Sneps-Sneppe, “A survey of smart cards data mining.,” in *AIST (Supplement)*, pp. 314–325, 2017.

[8] V. Venkatasubramanian, R. Rengaswamy, S. N. Kavuri, and K. Yin, “A review of process fault detection and diagnosis: Part iii: Process history based methods,” *Computers & chemical engineering*, vol. 27, no. 3, pp. 327–346, 2003.

[9] C. Aldrich and L. Auret, *Unsupervised process monitoring and fault diagnosis with machine learning methods.* Springer, 2013.

[10] N. Md Nor, C. R. Che Hassan, and M. A. Hussain, “A review of data-driven fault detection and diagnosis methods: Applications in chemical process systems,” *Reviews in Chemical Engineering*, vol. 36, no. 4, pp. 513–553, 2020.

[11] R. Isermann, *Fault-diagnosis applications: model-based condition monitoring: actuators, drives, machinery, plants, sensors, and fault-tolerant systems.* Springer Science & Business Media, 2011.

[12] M. Quiñones-Grueiro, A. Prieto-Moreno, C. Verde, and O. Llanes-Santiago, "Data-driven monitoring of multimode continuous processes: A review," *Chemometrics and Intelligent Laboratory Systems*, vol. 189, pp. 56–71, 2019.

[13] V. Venkatasubramanian, R. Rengaswamy, K. Yin, and S. N. Kavuri, "A review of process fault detection and diagnosis: Part i: Quantitative model-based methods," *Computers & chemical engineering*, vol. 27, no. 3, pp. 293–311, 2003.

[14] V. Venkatasubramanian, R. Rengaswamy, and S. N. Kavuri, "A review of process fault detection and diagnosis: Part ii: Qualitative models and search strategies," *Computers & chemical engineering*, vol. 27, no. 3, pp. 313–326, 2003.

[15] S. Qin, "Survey on data-driven industrial process monitoring and diagnosis," *Annual reviews in control*, vol. 36, no. 2, pp. 220–234, 2012.

[16] J. Hoskins, K. Kaliyur, and D. M. Himmelblau, "Fault diagnosis in complex chemical plants using artificial neural networks," *AIChE Journal*, vol. 37, no. 1, pp. 137–141, 1991.

[17] L. H. Chiang, M. E. Kotanchek, and A. K. Kordon, "Fault diagnosis based on fisher discriminant analysis and support vector machines," *Computers & chemical engineering*, vol. 28, no. 8, pp. 1389–1401, 2004.

[18] M. C. Thomas, W. Zhu, and J. A. Romagnoli, "Data mining and clustering in chemical process databases for monitoring and knowledge discovery," *Journal of Process Control*, vol. 67, pp. 160–175, 2018.

[19] W. Ku, R. H. Storer, and C. Georgakis, "Disturbance detection and isolation by dynamic principal component analysis," *Chemometrics and intelligent laboratory systems*, vol. 30, no. 1, pp. 179–196, 1995.

[20] M. Espadoto, R. M. Martins, A. Kerren, N. S. T. Hirata, and A. C. Telea, "Towards a quantitative survey of dimension reduction techniques," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2019.

[21] R. Caruana, M. Elhawary, N. Nguyen, and C. Smith, "Meta clustering," in *Sixth International Conference on Data Mining (ICDM'06)*, pp. 107–118, IEEE, 2006.

[22] B. Tang, M. Shepherd, E. Milios, and M. I. Heywood, "Comparing and combining dimension reduction techniques for efficient text clustering," in *Proceeding of SIAM international workshop on feature selection for data mining*, pp. 17–26, Citeseer, 2005.

[23] A. C. Belkina, C. O. Ciccolella, R. Anno, R. Halpert, J. Spidlen, and J. E. Snyder-Cappione, “Automated optimized parameters for t-distributed stochastic neighbor embedding improve visualization and analysis of large datasets,” *Nature communications*, vol. 10, no. 1, pp. 1–12, 2019.

[24] S. Beaver and A. Palazoğlu, “A cluster aggregation scheme for ozone episode selection in the san francisco, ca bay area,” *Atmospheric Environment*, vol. 40, no. 4, pp. 713–725, 2006.

[25] E. Hancer and D. Karaboga, “A comprehensive survey of traditional, merge-split and evolutionary approaches proposed for determination of cluster number,” *Swarm and Evolutionary Computation*, vol. 32, pp. 49–67, 2017.

[26] Q. Zhao, *Cluster validity in clustering methods.* PhD thesis, Itä-Suomen yliopisto, 2012.

[27] R. Tibshirani, G. Walther, and T. Hastie, “Estimating the number of clusters in a data set via the gap statistic,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 63, no. 2, pp. 411–423, 2001.

[28] D. Pelleg, A. W. Moore, *et al.*, “X-means: Extending k-means with efficient estimation of the number of clusters.,” in *Icml*, vol. 1, pp. 727–734, 2000.

[29] G. H. Ball and D. J. Hall, “Isodata, a novel method of data analysis and pattern classification,” tech. rep., Stanford research inst Menlo Park CA, 1965.

[30] S. Bandyopadhyay and U. Maulik, “Genetic clustering for automatic evolution of clusters and application to image classification,” *Pattern recognition*, vol. 35, no. 6, pp. 1197–1208, 2002.

[31] E. R. Hruschka, L. N. de Castro, and R. J. Campello, “Evolutionary algorithms for clustering gene-expression data,” in *Fourth IEEE International Conference on Data Mining (ICDM’04)*, pp. 403–406, IEEE, 2004.

[32] M. Mollaian, G. Dörgő, and A. Palazoglu, “Studying the synergy between dimension reduction and clustering methods to facilitate fault classification,” in *31st European Symposium on Computer Aided Process Engineering* (M. Türkay and R. Gani, eds.), vol. 50 of *Computer Aided Chemical Engineering*, pp. 819–824, Elsevier, 2021.

[33] M. Mollaian, G. Dörgő, and A. Palazoglu, “Performing multi-objective optimization and dimension reduction to detect the number of clusters.” Manuscript submitted for publication, N.D.

[34] J. F. Barragan, C. H. Fontes, and M. Embiruçu, “A wavelet-based clustering of multivariate time series using a multiscale spca approach,” *Computers & Industrial Engineering*, vol. 95, pp. 144–155, 2016.

[35] N. F. Thornhill, H. Melbø, and J. Wiik, "Multidimensional visualization and clustering of historical process data," *Industrial & engineering chemistry research*, vol. 45, no. 17, pp. 5971–5985, 2006.

[36] S. Gajjar, M. Kulahci, and A. Palazoglu, "Real-time fault detection and diagnosis using sparse principal component analysis," *Journal of Process Control*, vol. 67, pp. 112–128, 2018.

[37] A. Inselberg and B. Dimsdale, "Parallel coordinates: a tool for visualizing multi-dimensional geometry," in *Proceedings of the First IEEE Conference on Visualization: Visualization90*, pp. 361–378, IEEE, 1990.

[38] R. A. Becker, W. S. Cleveland, and M.-J. Shyu, "The visual design and control of trellis display," *Journal of computational and Graphical Statistics*, vol. 5, no. 2, pp. 123–155, 1996.

[39] T. Santhanam and M. Padmavathi, "Application of k-means and genetic algorithms for dimension reduction by integrating svm for diabetes diagnosis," *Procedia Computer Science*, vol. 47, pp. 76–83, 2015.

[40] M. Salehi, K. Aghilinasrollahabadi, and M. Salehi Esfandarani, "An investigation of stormwater quality variation within an industry sector using the self-reported data collected under the stormwater monitoring program," *Water*, vol. 12, no. 11, p. 3185, 2020.

[41] S. Han, S. Lee, and F. Peña-Mora, "Application of dimension reduction techniques for motion recognition: Construction worker behavior monitoring," in *Computing in Civil Engineering (2011)*, pp. 102–109, American Society of Civil Engineers, 2011.

[42] A. Cinar, A. Palazoglu, and F. Kayihan, *Chemical process performance evaluation.* CRC press, 2007.

[43] K. F. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[44] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[45] W. S. Torgerson, *Theory and methods of scaling.* Wiley, 1958.

[46] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[47] M. Balasubramanian, E. L. Schwartz, J. B. Tenenbaum, V. de Silva, and J. C. Langford, "The isomap algorithm and topological stability," *Science*, vol. 295, no. 5552, pp. 7–7, 2002.

[48] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[49] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[50] F. Gelgi, H. Davulcu, and S. Vadrevu, "Term ranking for clustering web search results.," in *WebDB*, Citeseer, 2007.

[51] O. Akman, T. Comar, D. Hrozencik, and J. Gonzales, "Data clustering and self-organizing maps in biology," in *Algebraic and Combinatorial Computational Biology*, pp. 351–374, Elsevier, 2019.

[52] R. Haralick and G. Kelly, "Pattern recognition with measurement space and spatial clustering for multiple images," *Proceedings of the IEEE*, vol. 57, no. 4, pp. 654–665, 1969.

[53] M. L. Zepeda-Mendoza and O. Resendis-Antonio, "Hierarchical agglomerative clustering," *Encyclopedia of systems biology*, vol. 43, no. 1, pp. 886–887, 2013.

[54] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, pp. 281–297, Oakland, CA, USA, 1967.

[55] H.-S. Park and C.-H. Jun, "A simple and fast algorithm for k-medoids clustering," *Expert systems with applications*, vol. 36, no. 2, pp. 3336–3341, 2009.

[56] C. E. Rasmussen *et al.*, "The infinite gaussian mixture model.," in *NIPS*, vol. 12, pp. 554–560, 1999.

[57] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise.," in *Kdd*, vol. 96, pp. 226–231, 1996.

[58] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: Ordering points to identify the clustering structure," *ACM Sigmod record*, vol. 28, no. 2, pp. 49–60, 1999.

[59] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *The Journal of Open Source Software*, vol. 2, mar 2017.

[60] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, "Automatic subspace clustering of high dimensional data," *Data Mining and Knowledge Discovery*, vol. 11, no. 1, pp. 5–33, 2005.

[61] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *The Journal of Machine Learning Research*, vol. 11, pp. 2837–2854, 2010.

[62] W. M. Rand, “Objective criteria for the evaluation of clustering methods,” *Journal of the American Statistical Association*, vol. 66, pp. 846–850, 1971.

[63] A. Rosenberg and J. Hirschberg, “V-measure: A conditional entropy-based external cluster evaluation measure,” in *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pp. 410–420, 2007.

[64] J. Palacio-Niño and F. Berzal, “Evaluation metrics for unsupervised learning algorithms,” *CoRR*, vol. abs/1905.05667, 2019.

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

[66] D. L. Davies and D. W. Bouldin, “A cluster separation measure,” *IEEE transactions on pattern analysis and machine intelligence*, no. 2, pp. 224–227, 1979.

[67] J. C. Dunn, “Well-separated clusters and optimal fuzzy partitions,” *Journal of cybernetics*, vol. 4, no. 1, pp. 95–104, 1974.

[68] T. Caliński and J. Harabasz, “A dendrite method for cluster analysis,” *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[69] D. Moulavi, P. A. Jaskowiak, R. J. Campello, A. Zimek, and J. Sander, “Density-based clustering validation,” in *Proceedings of the 2014 SIAM international conference on data mining*, pp. 839–847, SIAM, 2014.

[70] D. Xu and Y. Tian, “A comprehensive survey of clustering algorithms,” *Ann. Data. Sci.*, vol. 2, p. 165–193, 2015.

[71] J. J. Downs and E. F. Vogel, “A plant-wide industrial process control problem,” *Computers & chemical engineering*, vol. 17, no. 3, pp. 245–255, 1993.

[72] A. Bathelt, N. L. Ricker, and M. Jelali, “Revision of the Tennessee Eastman Process Model,” *IFAC-PapersOnLine*, vol. 48, pp. 309–314, 2015.

[73] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[74] S. Qin, "Statistical process monitoring: basics and beyond," *Journal of Chemometrics: A Journal of the Chemometrics Society*, vol. 17, no. 8-9, pp. 480–502, 2003.

[75] L. Ming and J. Zhao, "Review on chemical process fault detection and diagnosis," in *2017 6th International Symposium on Advanced Control of Industrial Processes (AdCONIP)*, pp. 457–462, 2017.

[76] S. Zheng and J. Zhao, "A new unsupervised data mining method based on the stacked autoencoder for chemical process fault diagnosis," *Computers & Chemical Engineering*, vol. 135, p. 106755, 2020.

[77] A. Zimmermann, "Method evaluation, parameterization, and result validation in unsupervised data mining: A critical survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 2, 2020.

[78] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.

[79] T. Van Craenendonck and H. Blockeel, "Using internal validity measures to compare clustering algorithms," *Benelearn 2015 Poster presentations (online)*, pp. 1–8, 2015.

[80] P. Ngatchou, A. Zarei, and A. El-Sharkawi, "Pareto multi objective optimization," in *Proceedings of the 13th International Conference on, Intelligent Systems Application to Power Systems*, pp. 84–91, 2005.

[81] S. Mishra, S. Saha, and S. Mondal, "Unsupervised method to ensemble results of multiple clustering solutions for bibliographic data," in *2017 IEEE Congress on Evolutionary Computation, CEC 2017 - Proceedings*, pp. 1459–1466, 2017.

[82] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "A survey of multiobjective evolutionary clustering," *ACM Computing Surveys*, vol. 47, no. 4, 2015.

[83] J. Handl and J. Knowles, "Exploiting the trade-off—the benefits of multiple objectives in data clustering," in *International Conference on Evolutionary Multi-Criterion Optimization*, pp. 547–560, Springer, 2005.

[84] S. Bandyopadhyay, U. Maulik, and A. Mukhopadhyay, "Multiobjective genetic clustering for pixel classification in remote sensing imagery," *IEEE transactions on Geoscience and Remote Sensing*, vol. 45, no. 5, pp. 1506–1511, 2007.

[85] M. T. Emmerich and A. H. Deutz, "A tutorial on multiobjective optimization: fundamentals and evolutionary methods," *Natural computing*, vol. 17, no. 3, pp. 585–609, 2018.

[86] K. Deb, “Multi-objective optimization,” in *Search methodologies*, pp. 403–449, Springer, 2014.

[87] D. Dua and C. Graff, “UCI machine learning repository,” 2017.

[88] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[89] V. M. Carrillo, O. Aguirre, and H. Taboada, “Applications and performance of the non-numerical ranking preferences method for post-pareto optimality,” *Procedia Computer Science*, vol. 6, pp. 243–248, 2011.

A Tennessee Eastman Process variables and faults

Table A.1: TEP manipulated variables

Variable name	Base case value (%)	Low limit	High limit	Units
D feed flow (stream 2)	63.053	0	5811	kgh^{-1}
E feed flow (stream 3)	53.980	0	8354	kgh^{-1}
A feed flow (stream 1)	24.644	0	1.017	kscmh
A and C feed flow (stream 4)	61.302	0	15.25	kscmh
Compressor recycle valve	22.210	0	100	%
Purge valve (stream 9)	40.064	0	100	%
Separator pot liquid flow (stream 10)	38.100	0	65.71	m^3h^{-1}
Stripper liquid product flow (stream 11)	46.534	0	49.10	m^3h^{-1}
Stripper steam valve	47.446	0	100	%
Reactor cooling water flow	41.106	0	227.1	m^3h^{-1}
Condenser cooling water flow	18.114	0	272.6	m^3h^{-1}
Agitator speed	50.000	150	250	rpm

Table A.2: TEP measured variables

Variable name	Base case value	Units
A feed (stream 1)	0.25052	kscmh
D feed (stream 2)	3664.0	kgh^{-1}
E feed (stream 3)	4509.3	kgh^{-1}
A and C feed (stream 4)	9.6477	kscmh
Recycle flow (stream 8)	26.902	kscmh
Reactor feed rate (stream 6)	42.339	kscmh
Reactor pressure	2705.0	kPa gauge
Reactor level	75.000	%
Reactor temperature	120.40	$^{\circ}C$
Purge rate (stream 9)	0.33712	kscmh
Product separator temperature	80.109	$^{\circ}C$
Product separator level	50.000	%
Product separator pressure	2633.7	kPa gauge
Product separator underflow (stream 10)	25.160	m^3h^{-1}
Stripper level	50.000	%
Stripper pressure	3102.2	kPa gauge
Stripper underflow (stream 11)	22.949	m^3h^{-1}
Stripper temperature	65.731	$^{\circ}C$
Stripper steam flow	230.31	kgh^{-1}
Compressor work	341.43	kW
Reactor cooling water outlet temperature	94.599	$^{\circ}C$
Separator cooling water outlet temperature	77.297	$^{\circ}C$

Table A.3: TEP faults

Fault number	Process variable	Type
IDV (1)	A/C feed ratio, B composition constant (stream 4)	Step
IDV (2)	B composition, A/C ratio constant (stream 4)	Step
IDV (3)	D feed temperature (stream 2)	Step
IDV (4)	Reactor cooling water inlet temperature	Step
IDV (5)	Condenser cooling water inlet temperature	Step
IDV (6)	A feed loss (stream 1)	Step
IDV (7)	C header pressure loss - reduced availability (stream 4)	Step
IDV (8)	A, B, C feed composition (stream 4)	Random variation
IDV (9)	D feed temperature (stream 2)	Random variation
IDV (10)	C feed temperature (stream 4)	Random variation
IDV (11)	Reactor cooling water inlet temperature	Random variation
IDV (12)	Condenser cooling water inlet temperature	Random variation
IDV (13)	Reaction kinetics	Slow drift
IDV (14)	Reactor cooling water valve	Sticking
IDV (15)	Condenser cooling water valve	Sticking
IDV (16)	Unknown	Unknown
IDV (17)	Unknown	Unknown
IDV (18)	Unknown	Unknown
IDV (19)	Unknown	Unknown
IDV (20)	Unknown	Unknown

B Codes

B.1 Code to demonstrate utilized methods

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
# %matplotlib inline
from sklearn.preprocessing import scale
import io

from sklearn.decomposition import PCA
from sklearn.manifold import Isomap
from sklearn.manifold import MDS
from sklearn.manifold import TSNE
from sklearn.manifold import LocallyLinearEmbedding


import sklearn.cluster as cluster
from scipy.cluster.hierarchy import dendrogram
from sklearn.datasets import make_blobs
from sklearn import mixture

import time
import sklearn.metrics as metrics
from sklearn.metrics import davies_bouldin_score

"""# prep"""

np.random.seed(0)
centers = [[7,0,-1], [10,3,-2], [6,-1,1]]
cluststd=[1,.5,.5]
df, t = make_blobs(n_samples=12, centers=centers,n_features=3, cluster_std=cluststd)

import plotly.express as px
import plotly.graph_objects as go
fig = px.scatter_3d(df, x=df[:,0], y=df[:,1], z=df[:,2], color=t)
fig.update_traces(marker={'size': 6})
fig.update_layout(
    scene=dict(
        xaxis=dict(showticklabels=False),
        yaxis=dict(showticklabels=False),
        zaxis=dict(showticklabels=False),
    ))
fig.show()

"""# PCA"""

pca = PCA(n_components=2)
pca_2d = pd.DataFrame(pca.fit_transform(df))
np.sum(pca.explained_variance_ratio_.round(2))

fig, ax = plt.subplots(sharex=False, figsize=(9,7))
plt.scatter(pca_2d[0][:], pca_2d[1][:], s=80, c=t, cmap=’plasma’)
plt.xticks(())
plt.yticks(())

"""# ISO"""

iso = Isomap(n_components=2, n_neighbors=3)
iso_2d = pd.DataFrame(iso.fit_transform(df))

fig, ax = plt.subplots(sharex=False, figsize=(9,7))
plt.scatter(iso_2d[0][:], iso_2d[1][:],cmap=’plasma’, s=80, c=t)
plt.xticks(())
plt.yticks(())

"""# MDS"""

mds = MDS(n_components=2,metric=True)
mds_2d = pd.DataFrame(mds.fit_transform(df))

fig, ax = plt.subplots(sharex=False, figsize=(9,7))
plt.scatter(mds_2d[0][:], mds_2d[1][:],cmap=’plasma’, s=80, c=t)
plt.xticks(())
plt.yticks(())

"""# t-SNE"""

tsne = TSNE(n_components=2, perplexity=2)
tsn_2d = pd.DataFrame(tsne.fit_transform(df))
```

```
fig, ax = plt.subplots(sharex=False, figsize=(9,7))
plt.scatter(tsn_2d[0][:], tsn_2d[1][:],cmap='plasma', s=80, c=t)
plt.xticks(())
plt.yticks(())

"""# UMAP"""

!pip install umap-learn
import umap

um = umap.UMAP(n_components=2, n_neighbors=3)
um_2d = pd.DataFrame(um.fit_transform(df))

fig, ax = plt.subplots(sharex=False, figsize=(9,7))
plt.scatter(um_2d[0][:], um_2d[1][:],cmap='plasma', s=80, c=t)
plt.xticks(())
plt.yticks(())

"""# Clust prep"""

def plot_clusters(data, algorithm, args, kwds):
    start_time = time.time()
    labels = algorithm(*args, **kwds).fit_predict(data)
    end_time = time.time()
    fig = px.scatter_3d(data, x=data[:,0], y=data[:,1], z=data[:,2], color=t, symbol =labels
                                                    , symbol_sequence= ['circle', 'x', 'square
                                                    ', 'circle-open'])
    fig.update_traces(marker={'size': 6})
    fig.update_layout(
        scene=dict(
        xaxis=dict(showticklabels=False),
        yaxis=dict(showticklabels=False),
        zaxis=dict(showticklabels=False),
    ))
    fig.show()
    return labels

# Commented out IPython magic to ensure Python compatibility.
def measure(data, labels, estimatorlabels_):
    print(82 * '_')
    print('homo\tcompl\tv-meas\tARI\tAMI\tsilhouette\tcalinski\tdb')
    print('%.3f\t%.3f\t%.3f\t%.3f\t%.3f\t%.3f\t\t%.3f\t%.3f'
#           % (metrics.homogeneity_score(labels, estimatorlabels_),
             metrics.completeness_score(labels, estimatorlabels_),
             metrics.v_measure_score(labels, estimatorlabels_),
             metrics.adjusted_rand_score(labels, estimatorlabels_),
             metrics.adjusted_mutual_info_score(labels,  estimatorlabels_),
             metrics.silhouette_score(data, estimatorlabels_,
                                      metric='euclidean'),
             metrics.calinski_harabasz_score(data, estimatorlabels_),
             davies_bouldin_score(data, estimatorlabels_)))

"""# Hiearchical"""

hward=plot_clusters(df, cluster.AgglomerativeClustering, (), {'n_clusters':3,'linkage':'
                                              average'})
measure(df,t, hward)

def plot_dendrogram(model, **kwargs):
    counts = np.zeros(model.children_.shape[0])
    n_samples = len(model.labels_)
    for i, merge in enumerate(model.children_):
        current_count = 0
        for child_idx in merge:
            if child_idx < n_samples:
                current_count += 1  # leaf node
            else:
                current_count += counts[child_idx - n_samples]
        counts[i] = current_count

    linkage_matrix = np.column_stack([model.children_, model.distances_,
                                      counts]).astype(float)
    # Plot the corresponding dendrogram
    dendrogram(linkage_matrix, **kwargs, color_threshold=0)

# setting distance_threshold=0 ensures we compute the full tree.
model = cluster.AgglomerativeClustering(distance_threshold=0, n_clusters=None)

model = model.fit(df)
plt.title('Hierarchical Clustering Dendrogram')
# plot the top three levels of the dendrogram
```

```
plt.plot(figsize=)
plot_dendrogram(model)
plt.xlabel("Index of point if no parenthesis")
plt.show()

"""# KMeans"""

mod = cluster.KMeans(n_clusters=3, random_state=0).fit(df)
labels=mod.labels_
fig = px.scatter_3d(df, x=df[:,0], y=df[:,1], z=df[:,2], color=t, symbol =labels,
                                              symbol_sequence= [’circle’, ’x’, ’square’])
fig.update_traces(marker={’size’: 6})
cent=mod.cluster_centers_
fig.add_traces(go.Scatter3d(x=cent[:,0],y=cent[:,1],z=cent[:,2],mode=’markers’, marker=dict(
                                              size=4,color=’black’,symbol=[’circle’, ’square
                                              ’, ’x’])))
fig.update_layout(
        scene=dict(
        xaxis=dict(showticklabels=False),
        yaxis=dict(showticklabels=False),
        zaxis=dict(showticklabels=False),
    ))
fig.show()
measure(df,t, labels)

"""# KMedoids"""

!pip install scikit-learn-extra
from sklearn_extra.cluster import KMedoids

mod = KMedoids(n_clusters=3, random_state=0).fit(df)
labels=mod.labels_
fig = px.scatter_3d(df, x=df[:,0], y=df[:,1], z=df[:,2], color=t, symbol =labels,
                                              symbol_sequence= [’circle’, ’x’, ’square’])
fig.update_traces(marker={’size’: 6})
cent=mod.cluster_centers_
fig.add_traces(go.Scatter3d(x=cent[:,0],y=cent[:,1],z=cent[:,2],mode=’markers’, marker=dict(
                                              size=3,color=’black’,symbol=[’circle’, ’square
                                              ’, ’x’])))
fig.update_layout(
        scene=dict(
        xaxis=dict(showticklabels=False),
        yaxis=dict(showticklabels=False),
        zaxis=dict(showticklabels=False),
    ))
fig.show()
measure(df,t, labels)

"""# dbscan grid search"""

def s1(est,X, y=t):
    model = est.fit(X)
    labels=model.labels_
    score = metrics.adjusted_rand_score(labels,t)
    return score
scoring= {'s1': s1}

from sklearn.model_selection import GridSearchCV
param_grid = {"min_samples": [1,2,3,4,5],'eps': [.2,.4,.6,.8,1,1.2,1.4,1.6]}
cv = [(slice(None), slice(None))]
search = GridSearchCV(
    cluster.DBSCAN(),
    param_grid,
    scoring=scoring,
    cv= cv, refit=False)

search.fit(df)

r=pd.DataFrame(search.cv_results_)
colors=['red','blue', 'green', 'yellow', 'purple', 'orange', 'black', 'lavender', 'aqua', '
                                              lawngreen', 'crimson', 'silver', 'bisque']
e=[.2,.4,.6,.8,1,1.2,1.4,1.6]
s=[1,2,3,4,5]
for er,m in zip(e,colors):
    es=r[r['param_eps']==er]
    y=es['mean_test_s1']
    x=es['param_min_samples']
    plt.plot(x,y,c=m)
    plt.xlabel("min_samples")
    plt.plot([], [], c=m, label='eps= %.1f'%(er))
    plt.legend(bbox_to_anchor=(1.05, 1))
plt.show()
```

```
"""# DBSCAN"""

db=plot_clusters(df, cluster.DBSCAN, (), {’min_samples’:4, ’eps’:1.6}) #eps
measure(df,t, db)

"""# optics grid search"""

from sklearn.model_selection import GridSearchCV
param_grid = {"min_samples": [1,2,3,4,5]}
cv = [(slice(None), slice(None))]
search = GridSearchCV(
    cluster.OPTICS(),
    param_grid,
    scoring=scoring,
    cv= cv, refit=False)
search.fit(df)

r=pd.DataFrame(search.cv_results_)
colors=[’red’,’blue’, ’green’, ’yellow’, ’purple’, ’orange’, ’black’, ’lavender’, ’aqua’, ’
                                              lawngreen’, ’crimson’, ’silver’, ’bisque’]
s=[1,2,3,4,5]
y=r[’mean_test_s1’]
x=r[’param_min_samples’]
plt.plot(x,y,c=m)
plt.show()
plt.plot

"""# OPTICS"""

op=plot_clusters(df, cluster.OPTICS, (), {’min_samples’:2})
measure(df,t, op)

"""# GMM"""

def s2(est,X, y=t):
    labels = est.fit_predict(X)
    score = metrics.adjusted_rand_score(labels,t)
    return score
scoring2= {’s2’: s2}
```

```
param_grid = {"n_components": [1,2,3,4,5],’covariance_type’:[’full’, ’tied’, ’diag’, ’spherical’]}
cv = [(slice(None), slice(None))]
search = GridSearchCV(
    mixture.GaussianMixture(),
    param_grid,
    scoring=scoring2,
    cv= cv, refit=False)
search.fit(df)
r=pd.DataFrame(search.cv_results_)
colors=[’red’,’blue’, ’green’, ’yellow’, ’purple’, ’orange’, ’black’, ’lavender’, ’aqua’, ’lawngreen’, ’crimson’, ’silver’, ’bisque’]
e=[’full’, ’tied’, ’diag’, ’spherical’]
s=[1,2,3,4,5]
for er,m in zip(e,colors):
    es=r[r[’param_covariance_type’]==er]
    y=es[’mean_test_s2’]
    x=es[’param_n_components’]
    plt.plot(x,y,c=m)
    plt.plot([], [], c=m, label=er)
    plt.legend(bbox_to_anchor=(1.05, 1))
plt.show()

gmm=plot_clusters(df, mixture.GaussianMixture, (), {’n_components’:3, ’covariance_type’:’full’})
measure(df,t, gmm)

"""#HDBSCAN"""

!pip install hdbscan
import hdbscan

param_grid = {"min_cluster_size": [2,3,4,5,6], ’min_samples’:[1,2,3,4,5,6]}
cv = [(slice(None), slice(None))]
search = GridSearchCV(
    hdbscan.HDBSCAN(),
    param_grid,
    scoring=scoring,
    cv= cv, refit=False)
search.fit(df)
```

```
r=pd.DataFrame(search.cv_results_)
colors=[’red’,’blue’, ’green’, ’yellow’, ’purple’, ’orange’, ’black’, ’lavender’, ’aqua’, ’
                                              lawngreen’, ’crimson’, ’silver’, ’bisque’]
e=[1,2,3,4,5,6]
s=[1,2,3,4,5]
for er,m in zip(e,colors):
    es=r[r[’param_min_samples’]==er]
    y=es[’mean_test_s1’]
    x=es[’param_min_cluster_size’]
    plt.plot(x,y,c=m)
    plt.xlabel("min_cs")
    plt.plot([], [], c=m, label=’min_samples= %.1f’%(er))
    plt.legend(bbox_to_anchor=(1.05, 1))
plt.show()

hd=plot_clusters(df, hdbscan.HDBSCAN, (), {’min_cluster_size’:3}) #eps
measure(df,t, hd)
hd.max()-hd.min()+1

"""#CLIQUE"""

! pip3 install pyclustering
from pyclustering.cluster.clique import clique, clique_visualizer

dat=np.array(df)
i=[2,3,4,5,6,7,8,9,10]
th=[0,1,2,3,4,5,6]
for ii in i:
  for tt in th:
    clique_instance = clique(dat, ii, tt)
    clique_instance.process()# start clustering process and obtain results
    clusters = clique_instance.get_clusters()  # allocated clusters
    cl=np.zeros((1, len(dat)))
    for i in range(len(clusters)):
      for j in clusters[i]:
         cl[0][j]=i
    print(’ii"’,ii, ’\ntt:’,tt)
    print(’ami’,(metrics.adjusted_mutual_info_score(t,cl[0])))
    print(’vm’,(metrics.v_measure_score(t, cl[0])))
    #print(’sc’,(metrics.silhouette_score(dat,cl[0])))
    #print(’db’,(davies_bouldin_score(dat, cl[0])))
    print('------------------------')

clique_instance = clique(df, 5, 0)
clique_instance.process()# start clustering process and obtain results
clusters = clique_instance.get_clusters()  # allocated clusters
cl=np.zeros((1, len(df)))
for i in range(len(clusters)):
  for j in clusters[i]:
     cl[0][j]=i

cl[0]

fig = px.scatter_3d(df, x=df[:,0], y=df[:,1], z=df[:,2], color=t, symbol =cl[0],
                                              symbol_sequence= ['circle', 'x', 'square', '
                                              circle-open'])
fig.update_traces(marker={'size': 6})
fig.update_layout(
        scene=dict(
        xaxis=dict(showticklabels=False),
        yaxis=dict(showticklabels=False),
        zaxis=dict(showticklabels=False),
    ))
fig.show()
```

B.2 Code to study the synergy between dimension reduction and clustering

```

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
# %matplotlib inline
from sklearn.preprocessing import scale
import io
from sklearn.decomposition import PCA
from sklearn.manifold import Isomap
from sklearn.manifold import MDS
from sklearn.manifold import TSNE
from sklearn.manifold import LocallyLinearEmbedding
import sklearn.cluster as cluster
from scipy.cluster.hierarchy import dendrogram
from sklearn import mixture
import time
import sklearn.metrics as metrics
from sklearn.metrics import davies_bouldin_score

"""# Data preparation"""

from google.colab import files
uploaded = files.upload() #or anyother method to upload

df = pd.read_excel(io.BytesIO(uploaded[’TEP_dataset_4states.xlsx’]), header=None)
truelabels=pd.read_excel(io.BytesIO(uploaded[’TEP_dataset_4states_labels.xlsx’]), header=
                                              None)
df=df.drop(df.apply(lambda x: min(x)==0 & max(x)==0, axis=1))
df=pd.DataFrame(scale(df))

"""# PCA"""

pca = PCA(n_components=2)
pca_2d = pca.fit_transform(df)

"""# MDS"""

mds = MDS(n_components=2,metric=True)
mds_2d=mds.fit_transform(df)

"""# ISO"""

iso = Isomap(n_components=2, n_neighbors=200)
iso_2d = pd.DataFrame(iso.fit_transform(df))

"""# t-SNE"""

tsne = TSNE(n_components=2, perplexity=100, random_state=12)
tsn_2d = pd.DataFrame(tsne.fit_transform(df))

"""# UMAP"""

!pip install umap-learn
import umap
```

```
um = umap.UMAP(n_components=2, n_neighbors=500,min_dist=5, spread=5, metric=’euclidean’)
um_2d = pd.DataFrame(um.fit_transform(df))

"""# Clustering preparation"""

def plot_clusters(data, algorithm, args, kwds):
    DR=’-’
    start_time = time.time()
    labels = algorithm(*args, **kwds).fit_predict(data)
    end_time = time.time()
    #palette = sns.color_palette(’bright’, np.unique(labels).max() + 1)
    #colors = [palette[x] if x >= 0 else (0.0, 0.0, 0.0) for x in labels]
    plt.figure(figsize=(7,5))
    plt.scatter(data[0], data[1], c=labels, cmap=’jet’,s=8)
    frame = plt.gca()
    frame.axes.get_xaxis().set_visible(False)
    frame.axes.get_yaxis().set_visible(False)
    plt.title(’Clusters found by {}\n Data: {}’.format(str(algorithm.__name__), str(DR)),
                                                  fontsize=24)
    #plt.text(’Clustering took {:.2f} s’.format(end_time - start_time), fontsize=20,
                                                  verticalalignment=’best’,)
    return labels

def measure(data, labels, estimatorlabels_):
    print(82 * ’_’)
    print(’homo\tcompl\tv-meas\tARI\tAMI\tsilhouette\tcalinski\tdb’)
    print(’%.3f\t%.3f\t%.3f\t%.3f\t%.3f\t%.3f\t\t%.3f\t%.3f’
#           % (metrics.homogeneity_score(labels, estimatorlabels_),
             metrics.completeness_score(labels, estimatorlabels_),
             metrics.v_measure_score(labels, estimatorlabels_),
             metrics.adjusted_rand_score(labels, estimatorlabels_),
             metrics.adjusted_mutual_info_score(labels,  estimatorlabels_),
             metrics.silhouette_score(data, estimatorlabels_,
                                      metric=’euclidean’),
             metrics.calinski_harabasz_score(data, estimatorlabels_),
             davies_bouldin_score(data, estimatorlabels_)))

"""# parameter search """

def s1(est,X, y=truelabels):
    labels=est.fit_predict(X)
    score=metrics.v_measure_score(truelabels, labels)
    return score
def s2(est,X, y=truelabels):
    labels=est.fit_predict(X)
    score=metrics.adjusted_rand_score((truelabels, labels)
    return score
def s3(est,X, y=truelabels):
    labels=est.fit_predict(X)
    score=metrics.adjusted_mutual_info_score((truelabels, labels)
    return score
scoring= {'s1': s1, 's2':s2, 's3': s3}

from sklearn.model_selection import GridSearchCV
DR=['pca', 'mds', 'iso', 'tsn', 'um']

"""#each method search"""
param_grid = {"n_clusters": [2,3,4,5,6,7],'linkage': ['ward','single','complete','average']}
cv = [(slice(None), slice(None))]
search = GridSearchCV(
    cluster.AgglomerativeClustering(),
    param_grid,
    scoring=scoring,
    cv= cv, refit=False)
heirarchical_result=[]
for i in DR:
  method=('{}_2d'.format(i))
  search.fit(method)
  hierarchical_result.append(pd.DataFrame(search.cv_results_))

for i in range(len(hierarchical_result)):
    print('hierarchical', 'method',DR[i],'\n', hierarchical_result[i].loc[
                                                hierarchical_result[i].mean_test_s1==
                                                hierarchical_result[i].mean_test_s1.max(),'\
                                                n', hierarchical_result[i].loc[
                                                hierarchical_result[i].mean_test_s2==
                                                hierarchical_result[i].mean_test_s2.max(), '
                                                \n', hierarchical_result[i].loc[
                                                hierarchical_result[i].mean_test_s3==
                                                hierarchical_result[i].mean_test_s3.max())
```

```
param_grid = {"n_clusters": [2,3,4,5,6,7,8]}
cv = [(slice(None), slice(None))]
search = GridSearchCV(
    cluster.KMeans(),
    param_grid,
    scoring=scoring,
    cv= cv, refit=False)

kmeans_result=[]
for i in DR:
  method=('{}_2d'.format(i))
  search.fit(method)
  kmeans_result.append(pd.DataFrame(search.cv_results_))

for i in range(len(kmeans_result)):
    print('kmeans', 'method',DR[i],'\n', kmeans_result[i].loc[kmeans_result[i].mean_test_s1=
                                                =kmeans_result[i].mean_test_s1.max(),'\n',
                                                kmeans_result[i].loc[kmeans_result[i].
                                                mean_test_s2==kmeans_result[i].mean_test_s2.
                                                max(), '\n', kmeans_result[i].loc[
                                                kmeans_result[i].mean_test_s3==kmeans_result
                                                [i].mean_test_s3.max())



!pip install scikit-learn-extra
from sklearn_extra.cluster import KMedoids

param_grid = {"n_clusters": [2,3,4,5,6,7,8]}
cv = [(slice(None), slice(None))]
search = GridSearchCV(
    KMedoids(),
    param_grid,
    scoring=scoring,
    cv= cv, refit=False)

kmedoids_result=[]
for i in DR:
  method=('{}_2d'.format(i))
  search.fit(method)
  kmedoids_result.append(pd.DataFrame(search.cv_results_))
```

```
for i in range(len(kmedoids_result)):
    print(’kmedoids’,’method’,DR[i],’\n’, kmedoids_result[i].loc[kmedoids_result[i].
                                                mean_test_s1==kmedoids_result[i].
                                                mean_test_s1.max(),’\n’, kmedoids_result[i].
                                                loc[kmedoids_result[i].mean_test_s2==
                                                kmedoids_result[i].mean_test_s2.max(), ’\n’,
                                                 kmedoids_result[i].loc[kmedoids_result[i].
                                                mean_test_s3==kmedoids_result[i].
                                                mean_test_s3.max())




param_grid = {"n_components": [2,3,4,5,6,7,8], "covariance_type":[’full’, ’tied’, ’diag’, ’
                                              spherical’]}
cv = [(slice(None), slice(None))]
search = GridSearchCV(
    mixture.GaussianMixture(),
    param_grid,
    scoring=scoring,
    cv= cv, refit=False)

gmm_result=[]
for i in DR:
  method=(’{}_2d’.format(i))
  search.fit(method)
  gmm_result.append(pd.DataFrame(search.cv_results_))

for i in range(len(gmm_result)):
    print(’gmm’,’method’,DR[i],’\n’, gmm_result[i].loc[gmm_result[i].mean_test_s1==
                                                gmm_result[i].mean_test_s1.max(),’\n’,
                                                gmm_result[i].loc[gmm_result[i].mean_test_s2
                                                ==gmm_result[i].mean_test_s2.max(), ’\n’,
                                                gmm_result[i].loc[gmm_result[i].mean_test_s3
                                                ==gmm_result[i].mean_test_s3.max())



param_grid = {"min_samples": [100,200,300,400,500,600,700],’eps’: [.5,2,3.5,5,6.5,8]}
cv = [(slice(None), slice(None))]
search = GridSearchCV(
    cluster.DBSCAN(),
    param_grid,
    scoring=scoring,
    cv= cv, refit=False)

dbs_result=[]
for i in DR:
  method=(’{}_2d’.format(i))
  search.fit(method)
  dbs_result.append(pd.DataFrame(search.cv_results_))

for i in range(len(dbs_result)):
    print(’dbs’,’method’,DR[i],’\n’, dbs_result[i].loc[dbs_result[i].mean_test_s1==
                                                dbs_result[i].mean_test_s1.max(),’\n’,
                                                dbs_result[i].loc[dbs_result[i].mean_test_s2
                                                ==dbs_result[i].mean_test_s2.max(), ’\n’,
                                                dbs_result[i].loc[dbs_result[i].mean_test_s3
                                                ==dbs_result[i].mean_test_s3.max())



param_grid = {"min_samples": [100,200,300,400,500,600],’max_eps’: [.5,2,3.5,6.5,8]}
cv = [(slice(None), slice(None))]
search = GridSearchCV(
    cluster.OPTICS(),
    param_grid,
    scoring=scoring,
    cv= cv, refit=False)

dbs_result=[]
for i in DR:
  method=(’{}_2d’.format(i))
  search.fit(method)
  ops_result.append(pd.DataFrame(search.cv_results_))

for i in range(len(ops_result)):
    print(’ops’,’method’,DR[i],’\n’, ops_result[i].loc[ops_result[i].mean_test_s1==
                                                ops_result[i].mean_test_s1.max(),’\n’,
                                                ops_result[i].loc[ops_result[i].mean_test_s2
                                                ==ops_result[i].mean_test_s2.max(), ’\n’,
                                                ops_result[i].loc[ops_result[i].mean_test_s3
                                                ==ops_result[i].mean_test_s3.max())
```

```
! pip3 install pyclustering
from pyclustering.cluster.clique import clique, clique_visualizer

intervals=[5,10,15,20,25,30,35,40]
threshold=[0,1,2,3,4,5,7]
for i in DR:
  method=('{}_2d'.format(i))
  for ii in intervals:
    sc_1=0
    sc_2=0
    sc_3=0
    for tt in thresholds:
      clique_instance = clique(method, ii, tt)
      clique_instance.process()# start clustering process and obtain results
      clusters = clique_instance.get_clusters()  # allocated clusters
      cl=np.zeros((1, len(df)))
      for iii in range(len(clusters)):
        for j in clusters[iii]:
           cl[0][j]=iii
      if metrics.adjusted_mutual_info_score(cl[0],truelabels)> sc_1:
   sc_1=metrics.adjusted_mutual_info_score(cl[0],truelabels)
      if metrics.adjusted_rand_score(cl[0],truelabels)> sc_2:
   sc_2=metrics.adjusted_rnad_score(cl[0],truelabels)
      if metrics.v_measure_score(cl[0],truelabels)> sc_3:
   sc_3=metrics.v_measure_score(cl[0],truelabels)
  for s in range(1,4):
    print('clique','method',DR[i],'\n', 'max_ami':, sc_1,'max_ari':,'\n', sc_2,'\n', 'max_v'
                                                  :, sc_3)
```

B.3 Code to find the number of clusters using multi-objective optimization

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
# %matplotlib inline
from sklearn.preprocessing import scale
import io
from sklearn.decomposition import PCA
from sklearn.manifold import Isomap
from sklearn.manifold import MDS
from sklearn.manifold import TSNE
from sklearn.manifold import LocallyLinearEmbedding
import sklearn.cluster as cluster
from scipy.cluster.hierarchy import dendrogram
from sklearn import mixture
import time
import sklearn.metrics as metrics
from sklearn.metrics import davies_bouldin_score
from sklearn.datasets import make_blobs
from sklearn import datasets
from sklearn.model_selection import GridSearchCV
!pip install umap-learn
import umap
!pip install scikit-learn-extra
from sklearn_extra.cluster import KMedoids



"""# Data Preparation"""


 #to create the synthetic dataset
np.random.seed(0)
centers = [[.5, 2], [-1.5, 0], [1, -1], [-4,-3],[-1.7,-5]]
n_clusters = len(centers)
df, truelabels = make_blobs(n_samples=3000, centers=centers, cluster_std=0.6)
plt.scatter(df[:,0], df[:,1], c=truelabels, alpha=.5)
plt.title('Synthetic data')
plt.xticks(())
plt.yticks(())

#to upload any other dataset: wine
from google.colab import files
uploaded = files.upload() #or anyother uploading method

df = pd.read_csv(io.BytesIO(uploaded['wine.data']), header=None)
truelabels = df[0]
df=df.iloc[:,1:]
df=scale(df)

pd.DataFrame(t).value_counts()
```

```
#to upload any other dataset: TEP
df = pd.read_excel(io.BytesIO(uploaded[’TEP_dataset_4states.xlsx’]), header=None)
truelabels=pd.read_excel(io.BytesIO(uploaded[’TEP_dataset_4states_labels.xlsx’]), header=
                                              None)
df=df.drop(df.apply(lambda x: min(x)==0 & max(x)==0, axis=1))
df=pd.DataFrame(scale(df))

"""# DBCV initialization"""

!pip install hdbscan
import hdbscan

import numpy as np
from sklearn.metrics import pairwise_distances
from scipy.spatial.distance import cdist
from hdbscan._hdbscan_linkage import mst_linkage_core
from hdbscan.hdbscan_ import isclose

def all_points_core_distance(distance_matrix, d=2.0):
    """
    Compute the all-points-core-distance for all the points of a cluster.

    Parameters
    ----------
    distance_matrix : array (cluster_size, cluster_size)
        The pairwise distance matrix between points in the cluster.

    d : integer
        The dimension of the dataset, which is used in the computation
        of the all-point-core-distance as per the paper.

    Returns
    -------
    core_distances : array (cluster_size,)
        The all-points-core-distance of each point in the cluster

    References
    ----------
    Moulavi, D., Jaskowiak, P.A., Campello, R.J., Zimek, A. and Sander, J.,
    2014. Density-Based Clustering Validation. In SDM (pp. 839-847).
    """
    distance_matrix[distance_matrix != 0] = (1.0 / distance_matrix[
        distance_matrix != 0]) ** d
    result = distance_matrix.sum(axis=1)
    result /= distance_matrix.shape[0] - 1
    result **= (-1.0 / d)

    return result


def all_points_mutual_reachability(X, labels, cluster_id,
                                   metric='euclidean', d=None, **kwd_args):
    """
    Compute the all-points-mutual-reachability distances for all the points of
    a cluster.

    If metric is 'precomputed' then assume X is a distance matrix for the full
    dataset. Note that in this case you must pass in 'd' the dimension of the
    dataset.

    Parameters
    ----------
    X : array (n_samples, n_features) or (n_samples, n_samples)
        The input data of the clustering. This can be the data, or, if
        metric is set to `precomputed` the pairwise distance matrix used
        for the clustering.

    labels : array (n_samples)
        The label array output by the clustering, providing an integral
        cluster label to each data point, with -1 for noise points.

    cluster_id : integer
        The cluster label for which to compute the all-points
        mutual-reachability (which should be done on a cluster
        by cluster basis).

    metric : string
        The metric used to compute distances for the clustering (and
        to be re-used in computing distances for mr distance). If
        set to `precomputed` then X is assumed to be the precomputed
        distance matrix between samples.
```

```
d : integer (or None)
    The number of features (dimension) of the dataset. This need only
    be set in the case of metric being set to `precomputed`, where
    the ambient dimension of the data is unknown to the function.

**kwd_args :
    Extra arguments to pass to the distance computation for other
    metrics, such as minkowski, Mahanalobis etc.

Returns
-------

mutual_reachaibility : array (n_samples, n_samples)
    The pairwise mutual reachability distances between all points in `X`
    with `label` equal to `cluster_id`.

core_distances : array (n_samples,)
    The all-points-core_distance of all points in `X` with `label` equal
    to `cluster_id`.

References
----------
Moulavi, D., Jaskowiak, P.A., Campello, R.J., Zimek, A. and Sander, J.,
2014. Density-Based Clustering Validation. In SDM (pp. 839-847).
"""
if metric == 'precomputed':
    if d is None:
        raise ValueError('If metric is precomputed a '
                         'd value must be provided!')
    distance_matrix = X[labels == cluster_id, :][:, labels == cluster_id]
else:
    subset_X = X[labels == cluster_id, :]
    distance_matrix = pairwise_distances(subset_X, metric=metric,
                                         **kwd_args)
    d = X.shape[1]

core_distances = all_points_core_distance(distance_matrix.copy(), d=d)
core_dist_matrix = np.tile(core_distances, (core_distances.shape[0], 1))

result = np.dstack(
        [distance_matrix, core_dist_matrix, core_dist_matrix.T]).max(axis=-1)

    return result, core_distances


def internal_minimum_spanning_tree(mr_distances):
    """
    Compute the 'internal' minimum spanning tree given a matrix of mutual
    reachability distances. Given a minimum spanning tree the 'internal'
    graph is the subgraph induced by vertices of degree greater than one.

    Parameters
    ----------
    mr_distances : array (cluster_size, cluster_size)
        The pairwise mutual reachability distances, inferred to be the edge
        weights of a complete graph. Since MSTs are computed per cluster
        this is the all-points-mutual-reacability for points within a single
        cluster.

    Returns
    -------
    internal_nodes : array
        An array listing the indices of the internal nodes of the MST

    internal_edges : array (?, 3)
        An array of internal edges in weighted edge list format; that is
        an edge is an array of length three listing the two vertices
        forming the edge and weight of the edge.

    References
    ----------
    Moulavi, D., Jaskowiak, P.A., Campello, R.J., Zimek, A. and Sander, J.,
    2014. Density-Based Clustering Validation. In SDM (pp. 839-847).
    """
    single_linkage_data = mst_linkage_core(mr_distances)
    min_span_tree = single_linkage_data.copy()
    for index, row in enumerate(min_span_tree[1:], 1):
        candidates = np.where(isclose(mr_distances[int(row[1])], row[2]))[0]
        candidates = np.intersect1d(candidates,
                                    single_linkage_data[:index, :2].astype(
                                        int))
        candidates = candidates[candidates != row[1]]
        assert len(candidates) > 0
        row[0] = candidates[0]

    vertices = np.arange(mr_distances.shape[0])[
        np.bincount(min_span_tree.T[:2].flatten().astype(np.intp)) > 1]
    # A little "fancy" we select from the flattened array reshape back
    # (Fortran format to get indexing right) and take the product to do an and
    # then convert back to boolean type.
    edge_selection = np.prod(np.in1d(min_span_tree.T[:2], vertices).reshape(
        (min_span_tree.shape[0], 2), order='F'), axis=1).astype(bool)

    # Density sparseness is not well defined if there are no
    # internal edges (as per the referenced paper). However
    # MATLAB code from the original authors simply selects the
    # largest of *all* the edges in the case that there are
    # no internal edges, so we do the same here
    if np.any(edge_selection):
        # If there are any internal edges, then subselect them out
        edges = min_span_tree[edge_selection]
    else:
        # If there are no internal edges then we want to take the
        # max over all the edges that exist in the MST, so we simply
        # do nothing and return all the edges in the MST.
        edges = min_span_tree.copy()

    return vertices, edges


def density_separation(X, labels, cluster_id1, cluster_id2,
                       internal_nodes1, internal_nodes2,
                       core_distances1, core_distances2,
                       metric='euclidean', **kwd_args):
    """
    Compute the density separation between two clusters. This is the minimum
    all-points mutual reachability distance between pairs of points, one from
    internal nodes of MSTs of each cluster.

    Parameters
    ----------
    X : array (n_samples, n_features) or (n_samples, n_samples)
```

```
        The input data of the clustering. This can be the data, or, if
        metric is set to ‘precomputed‘ the pairwise distance matrix used
        for the clustering.

    labels : array (n_samples)
        The label array output by the clustering, providing an integral
        cluster label to each data point, with -1 for noise points.

    cluster_id1 : integer
        The first cluster label to compute separation between.

    cluster_id2 : integer
        The second cluster label to compute separation between.

    internal_nodes1 : array
        The vertices of the MST for ‘cluster_id1‘ that were internal vertices.

    internal_nodes2 : array
        The vertices of the MST for ‘cluster_id2‘ that were internal vertices.

    core_distances1 : array (size of cluster_id1,)
        The all-points-core_distances of all points in the cluster
        specified by cluster_id1.

    core_distances2 : array (size of cluster_id2,)
        The all-points-core_distances of all points in the cluster
        specified by cluster_id2.

    metric : string
        The metric used to compute distances for the clustering (and
        to be re-used in computing distances for mr distance). If
        set to ‘precomputed‘ then X is assumed to be the precomputed
        distance matrix between samples.

    **kwd_args :
        Extra arguments to pass to the distance computation for other
        metrics, such as minkowski, Mahanalobis etc.

    Returns
    -------
    The ’density separation’ between the clusters specified by
    `cluster_id1` and `cluster_id2`.

    References
    ----------
    Moulavi, D., Jaskowiak, P.A., Campello, R.J., Zimek, A. and Sander, J.,
    2014. Density-Based Clustering Validation. In SDM (pp. 839-847).
    """
    if metric == 'precomputed':
        sub_select = X[labels == cluster_id1, :][:, labels == cluster_id2]
        distance_matrix = sub_select[internal_nodes1, :][:, internal_nodes2]
    else:
        cluster1 = X[labels == cluster_id1][internal_nodes1]
        cluster2 = X[labels == cluster_id2][internal_nodes2]
        distance_matrix = cdist(cluster1, cluster2, metric, **kwd_args)

    core_dist_matrix1 = np.tile(core_distances1[internal_nodes1],
                                (distance_matrix.shape[1], 1)).T
    core_dist_matrix2 = np.tile(core_distances2[internal_nodes2],
                                (distance_matrix.shape[0], 1))

    mr_dist_matrix = np.dstack([distance_matrix,
                                core_dist_matrix1,
                                core_dist_matrix2]).max(axis=-1)

    return mr_dist_matrix.min()


def validity_index(X, labels, metric='euclidean',
                   d=None, per_cluster_scores=False, **kwd_args):
    """
    Compute the density based cluster validity index for the
    clustering specified by `labels` and for each cluster in `labels`.

    Parameters
    ----------
    X : array (n_samples, n_features) or (n_samples, n_samples)
        The input data of the clustering. This can be the data, or, if
        metric is set to `precomputed` the pairwise distance matrix used
        for the clustering.

    labels : array (n_samples)
        The label array output by the clustering, providing an integral
        cluster label to each data point, with -1 for noise points.

    metric : optional, string (default 'euclidean')
        The metric used to compute distances for the clustering (and
        to be re-used in computing distances for mr distance). If
        set to `precomputed` then X is assumed to be the precomputed
        distance matrix between samples.

    d : optional, integer (or None) (default None)
        The number of features (dimension) of the dataset. This need only
        be set in the case of metric being set to `precomputed`, where
        the ambient dimension of the data is unknown to the function.

    per_cluster_scores : optional, boolean (default False)
        Whether to return the validity index for individual clusters.
        Defaults to False with the function returning a single float
        value for the whole clustering.

    **kwd_args :
        Extra arguments to pass to the distance computation for other
        metrics, such as minkowski, Mahanalobis etc.

    Returns
    -------
    validity_index : float
        The density based cluster validity index for the clustering. This
        is a numeric value between -1 and 1, with higher values indicating
        a 'better' clustering.

    per_cluster_validity_index : array (n_clusters,)
        The cluster validity index of each individual cluster as an array.
        The overall validity index is the weighted average of these values.
        Only returned if per_cluster_scores is set to True.

    References
    ----------
    Moulavi, D., Jaskowiak, P.A., Campello, R.J., Zimek, A. and Sander, J.,
    2014. Density-Based Clustering Validation. In SDM (pp. 839-847).
    """
    core_distances = {}
```

```
density_sparseness = {}
mst_nodes = {}
mst_edges = {}

max_cluster_id = labels.max() + 1
density_sep = np.inf * np.ones((max_cluster_id, max_cluster_id),
                               dtype=np.float64)
cluster_validity_indices = np.empty(max_cluster_id, dtype=np.float64)

for cluster_id in range(max_cluster_id):

    if np.sum(labels == cluster_id) == 0:
        continue

    mr_distances, core_distances[
        cluster_id] = all_points_mutual_reachability(
        X,
        labels,
        cluster_id,
        metric,
        d,
        **kwd_args
    )

    mst_nodes[cluster_id], mst_edges[cluster_id] = \
        internal_minimum_spanning_tree(mr_distances)
    density_sparseness[cluster_id] = mst_edges[cluster_id].T[2].max()

for i in range(max_cluster_id):

    if np.sum(labels == i) == 0:
        continue

    internal_nodes_i = mst_nodes[i]
    for j in range(i + 1, max_cluster_id):

        if np.sum(labels == j) == 0:
            continue

        internal_nodes_j = mst_nodes[j]
        density_sep[i, j] = density_separation(
                X, labels, i, j,
                internal_nodes_i, internal_nodes_j,
                core_distances[i], core_distances[j],
                metric=metric, **kwd_args
            )
            density_sep[j, i] = density_sep[i, j]

    n_samples = float(X.shape[0])
    result = 0

    for i in range(max_cluster_id):

        if np.sum(labels == i) == 0:
            continue

        min_density_sep = density_sep[i].min()
        cluster_validity_indices[i] = (
            (min_density_sep - density_sparseness[i]) /
            max(min_density_sep, density_sparseness[i])
        )
        cluster_size = np.sum(labels == i)
        result += (cluster_size / n_samples) * cluster_validity_indices[i]

    if per_cluster_scores:
        return result, cluster_validity_indices
    else:
        return result

"""# Pareto Optimization"""

def is_pareto_efficient(costs, return_mask = True):
    """
    Find the pareto-efficient points
    :param costs: An (n_points, n_costs) array
    :param return_mask: True to return a mask
    :return: An array of indices of pareto-efficient points.
        If return_mask is True, this will be an (n_points, ) boolean array
        Otherwise it will be a (n_efficient_points, ) integer array of indices.
    """
    is_efficient = np.arange(costs.shape[0])
    n_points = costs.shape[0]
    next_point_index = 0  # Next index in the is_efficient array to search for
    while next_point_index<len(costs):
        nondominated_point_mask = np.any(costs>costs[next_point_index], axis=1) #true false
                                                       mide
        nondominated_point_mask[next_point_index] = True
        is_efficient = is_efficient[nondominated_point_mask]  # Remove dominated points,
                                                       adade unai ke truan
        costs = costs[nondominated_point_mask] #khode unai ke true budn
        next_point_index = np.sum(nondominated_point_mask[:next_point_index])+1
    if return_mask:
        is_efficient_mask = np.zeros(n_points, dtype = bool)
        is_efficient_mask[is_efficient] = True
        return is_efficient_mask
    else:
        return is_efficient



"""# Create lower dimensional data"""

pca = PCA(n_components=2)
pca_2d = pca.fit_transform(df)

mds = MDS(n_components=2,metric=True, random_state=310)
mds_2d=mds.fit_transform(df)

iso = Isomap(n_components=2, n_neighbors=70)
iso_2d = iso.fit_transform(df)

tsne = TSNE(n_components=2, perplexity=100, random_state=12)
tsn_2d = np.array(tsne.fit_transform(df))
tsn_2d=tsn_2d.astype(’float64’)

um = umap.UMAP(n_components=2, n_neighbors=400,metric=’euclidean’, random_state=310,
                                               min_dist=2.5, spread=3, local_connectivity=1)
um_2d = um.fit_transform(df)
um_2d=um_2d.astype(’float64’)



"""# Search and optimization"""

def s1(est,X, y=truelabels): #ami
```

```
    labels=est.fit_predict(X)
    score = metrics.adjusted_mutual_info_score(labels,y)
    return score
def s2(est,X, y=truelabels): #vmeasure
    labels=est.fit_predict(X)
    score = metrics.v_measure_score(y, labels)
    return score
def s3(est,X, y=None): #silhouette
    labels=est.fit_predict(X)
    try:
      score = metrics.silhouette_score(X, labels)
    except:
      score=0
    return score
def s4(est,X, y=None): #db
    labels=est.fit_predict(X)
    try:
      score = davies_bouldin_score(X, labels)
      score=.1/score
    except:
      score=0
    return score
def s5(est,X, y=None): #dbcv
    labels=est.fit_predict(X)
    try:
      score = validity_index(X,labels)
    except:
      score=0
    return score
def s6 (est,X,y=None): #clust number for density based methods
    labels=est.fit_predict(X)
    score=labels.max()-labels.min()+1
    return score
scoring= {’ami’: s1,’v’: s2, ’sil’:s3, ’db’:s4, ’dbcv’:s5, ’count’:s6}

#select the parameter search space, depending on the clustering method

#param_grid = {"n_clusters": [2,3,4,5,6, 7,8,9,10],’linkage’: [’ward’,’single’,’complete’,’
                                              average’]} #heirarchical
#param_grid = {"n_clusters": [2,3,4,5,6,7,8,9,10,11]} #kmeans, kmedoids
```

```
#param_grid = {"n_components": [2,3,4,5,6,7,8, 9, 10,11], "covariance_type":[’full’, ’tied’,
                                               ’diag’, ’spherical’]} #gmm
param_grid = {"min_samples": [5,10,15,20,25,30,35,40,45,50],’eps’: [.1,.3,.5,.7,.9,1.1,1.3,1
                                              .5,1.7,1.9,2.1]} #dbscan
#param_grid = {"min_samples": [20,40,60,80,100],’max_eps’: [5,10,15,20,25]} #optics
#param_grid = {"min_samples": [2,5,10,15,20,30,40],’min_cluster_size’: [5,20,25,30,35]} #
                                              hdbscan

cv = [(slice(None), slice(None))]
search = GridSearchCV(
    cluster.DBSCAN( ),
    param_grid,
    scoring=scoring,
    cv= cv, refit=False)

new_clustering_solutions=[]

df_2d=df #for the sake of loop. DF is NOT 2d
DR=[’df’, ’pca’, ’mds’, ’iso’, ’tsn’, ’um’]
for i in DR:
  method=(’{}_2d’.format(i))
  search.fit(method)
  all_solutions=pd.DataFrame(search.cv_results_)

  all_needed=[’params’,’mean_test_ami’, ’mean_test_v’, ’mean_test_sil’, ’mean_test_db’, ’
                                                mean_test_dbcv’, ’mean_test_count’]
  overview=all_solutions[all_needed]
  objectives=[’mean_test_sil’,’mean_test_db’, ’mean_test_dbcv’]
  for_pareto=all_solutions[objectives]
  for_pareto[’mean_test_dbcv’]=for_pareto[’mean_test_dbcv’].fillna(0)
  nondominated = is_pareto_efficient(for_pareto.values, return_mask = True)

  overview[’pareto’]=nondominated
  threshold_check=[]
  for x in range(len(overview)):
    if overview.pareto.loc[x] and (abs(for_pareto.loc[x])>=0.01).all():
      threshold_check.append(’useful_solution’)
    elif ~overview.pareto.loc[x]:
      threshold_check.append(’not_solution’)
    else:
      threshold.append(’not_useful_solution’)
  overview[’threshold_check’]=threshold_check
  new_clustering_solutions.append(overview)


for (each_dr, i) in zip(new_clustering_solutions, DR):
  solution_count=pd.DataFrame(each_dr.groupby(’mean_test_count’)[’mean_test_count’].count())
  solution_count[’fraction’]=solution_count.mean_test_count/len(solution_count)
  solution_count.rename(columns={"mean_test_count": "how many"}, inplace=True)
  print(i, solution_count)


"""# plot a specific solution set"""

from matplotlib.ticker import StrMethodFormatter
x=overview[’mean_test_sil’]
y=overview[’mean_test_db’]
plt.rcParams["figure.figsize"] = (20,20)
markers = [’o’, ’x’]
vals=[False,True]
for marker, val in zip(markers, vals):
    toUse = overview.pareto == val
    plt.scatter(x[toUse],y[toUse], c=’k’, marker=marker, s=70)
plt.gca().set_aspect(’equal’, adjustable=’box’)
plt.axhline(y=0.01, color="black", linestyle="--")
plt.axhline(y=-0.01, color="black", linestyle="--")
plt.axvline(x=0.01, color="black", linestyle="--")
plt.axvline(x=-0.01, color="black", linestyle="--")
plt.xlabel(’Silhouette Cofficient’, fontsize=15)
plt.xticks(fontsize=15)
plt.ylabel(’Davies-Bouldin Index’, fontsize=15)
plt.gca().yaxis.set_major_formatter(StrMethodFormatter(’{x:,.2f}’))
plt.gca().xaxis.set_major_formatter(StrMethodFormatter(’{x:,.2f}’))
plt.yticks(fontsize=15)
plt.plot([], [],"o", label="Dominated solution", color=’black’)
plt.plot([],[], "x", label="Non-dominated solution", color=’black’)
plt.legend(fontsize=’large’)
plt.show()

plt.rcParams["figure.figsize"] = (10,3)
plt.plot([], [],"o", label="Dominated solution", color=’black’)
plt.plot([],[], "x", label="Nondominated solution, meeting the threshold", color=’black’)
plt.plot([],[], "x", label="Nondominated solution, not meeting the threshold", color=’red’)
plt.legend(fontsize=’large’, markerscale=1)
plt.show()



#plot 3d
import plotly.express as px
fig = px.scatter_3d(overview, x=’mean_test_sil’, y=’mean_test_db’, z=’mean_test_dbcv’,
              symbol = overview.pareto,
                 symbol_sequence= [’circle’, ’x’], color=overview.threshold_check,
                                                                color_discrete_sequence=[’
                                                                black’, ’black’, ’red’])
fig.update_traces(marker={’size’: 3})
fig.update_layout(
    scene = dict(
        xaxis = dict(nticks=12, range=[-0.2,0.6],),
                     yaxis = dict(nticks=8, range=[-0.1,0.3],),
                     zaxis = dict(nticks=10, range=[-0.2,0.45])))
fig.update_layout(scene = dict(
                    xaxis_title=’Silhouette Coefficient’,
                    yaxis_title=’Davies-Bouldin Index’,
                    zaxis_title=’DBCV Index’))

from itertools import cycle
names = cycle([’Dominated solution’, ’Nondominated solution, meeting the threshold’,  ’
                                              Nondominated solution, not meeting the
                                              threshold’])
fig.for_each_trace(lambda t:  t.update(name = next(names)))



fig.show()

import plotly.io as pio
bh=pio.write_html(fig, file=’index.html’, auto_open=True) #save the interactive 3d plot



"""# plotting numbers vs fractions"""

#for desired dataset
#for desired clustering method
labels=[’No DR’,’PCA’,’MDS’,’ISOMAP’,’t-SNE’,’UMAP’]
```

```
plt.rcParams["figure.figsize"] = (9,7)
for (each_dr , i) in zip(new_clustering_solutions , DR):
  solution_count=pd.DataFrame(each_dr.groupby(’mean_test_count’)[’mean_test_count’].count())
  solution_count[’fraction’]=solution_count.mean_test_count/len(solution_count)
  solution_count.rename(columns={"mean_test_count": "how_many"}, inplace=True)
  solution_count.reset_index(inplace=True)
  x1=np.arange(solution_count.mean_test_count.min(), solution_count.mean_test_count.max())
  y1=solution_count.fraction
  plt.plot(x1,y1, ’--’, linewidth=’2’, label=labels[i], marker=’o’, markersize=7)
  plt.xticks(fontsize=15)
  plt.yticks( fontsize=15)
  plt.legend()
  plt.xlabel(’Number of clusters’, fontsize=15)
  plt.ylabel(’Fraction of solutions’, fontsize=15)
```

C Results of Multi-Objective Clustering

Current section contains all the solutions found from studying clustering hyperparameters over their search space for Chapter 4. Each table is representing the results of a clustering method. The rows are presenting different solutions indicated with their respective parameters. The columns show all calculated scores (whether internal or external), number of found clusters, whether the solution is a Pareto solution or not and if it is outside the limiting threshold. The solution was used for finding the cluster number if both **Dominant** and **Threshold** values were *True*.

Table C.1: Wine dataset - Hierarchical clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’linkage’: ’ward’, ’n clusters’: 2	0.56	0.56	0.27	0.07	-0.33	2.00	True	True
’linkage’: ’ward’, ’n clusters’: 3	0.78	0.79	0.28	0.07	-0.28	3.00	True	True
’linkage’: ’ward’, ’n clusters’: 4	0.71	0.71	0.23	0.06	-0.28	4.00	False	False
’linkage’: ’ward’, ’n clusters’: 5	0.67	0.67	0.19	0.05	-0.32	5.00	False	False
’linkage’: ’ward’, ’n clusters’: 6	0.66	0.67	0.18	0.06	-0.31	6.00	False	False
’linkage’: ’ward’, ’n clusters’: 7	0.64	0.65	0.19	0.06	-0.26	7.00	False	False
’linkage’: ’ward’, ’n clusters’: 8	0.62	0.63	0.19	0.06	-0.27	8.00	False	False
’linkage’: ’ward’, ’n clusters’: 9	0.59	0.61	0.19	0.06	-0.26	9.00	False	False
’linkage’: ’ward’, ’n clusters’: 10	0.58	0.60	0.20	0.07	-0.25	10.00	False	False
’linkage’: ’single’, ’n clusters’: 2	0.01	0.03	0.22	0.10	0.18	2.00	True	True
’linkage’: ’single’, ’n clusters’: 3	0.01	0.03	0.18	0.11	0.00	3.00	False	False
’linkage’: ’single’, ’n clusters’: 4	0.01	0.04	0.18	0.12	0.00	4.00	False	False
’linkage’: ’single’, ’n clusters’: 5	0.01	0.05	0.14	0.13	0.00	5.00	False	False
’linkage’: ’single’, ’n clusters’: 6	0.01	0.06	0.12	0.13	0.00	6.00	False	False
’linkage’: ’single’, ’n clusters’: 7	0.01	0.06	0.05	0.13	0.00	7.00	False	False
’linkage’: ’single’, ’n clusters’: 8	0.02	0.08	0.02	0.13	0.00	8.00	False	False
’linkage’: ’single’, ’n clusters’: 9	0.01	0.08	0.02	0.14	0.00	9.00	False	False
’linkage’: ’single’, ’n clusters’: 10	0.01	0.09	-0.01	0.14	0.00	10.00	False	False
’linkage’: ’complete’, ’n clusters’: 2	0.35	0.35	0.16	0.05	-0.46	2.00	False	False
’linkage’: ’complete’, ’n clusters’: 3	0.61	0.61	0.20	0.05	-0.38	3.00	False	False
’linkage’: ’complete’, ’n clusters’: 4	0.66	0.67	0.19	0.06	-0.36	4.00	False	False
’linkage’: ’complete’, ’n clusters’: 5	0.70	0.71	0.19	0.06	-0.33	5.00	False	False
’linkage’: ’complete’, ’n clusters’: 6	0.69	0.69	0.18	0.07	0.00	6.00	False	False
’linkage’: ’complete’, ’n clusters’: 7	0.71	0.72	0.19	0.07	0.00	7.00	False	False
’linkage’: ’complete’, ’n clusters’: 8	0.70	0.71	0.19	0.07	0.00	8.00	False	False
’linkage’: ’complete’, ’n clusters’: 9	0.67	0.68	0.18	0.07	0.00	9.00	False	False
’linkage’: ’complete’, ’n clusters’: 10	0.63	0.64	0.17	0.07	0.00	10.00	False	False
’linkage’: ’average’, ’n clusters’: 2	-0.00	0.01	0.26	0.17	0.00	2.00	True	False
’linkage’: ’average’, ’n clusters’: 3	-0.00	0.02	0.16	0.10	0.00	3.00	False	False
’linkage’: ’average’, ’n clusters’: 4	0.02	0.05	0.15	0.09	0.00	4.00	False	False
’linkage’: ’average’, ’n clusters’: 5	0.53	0.54	0.23	0.09	0.00	5.00	False	False
’linkage’: ’average’, ’n clusters’: 6	0.52	0.53	0.21	0.09	0.00	6.00	False	False
’linkage’: ’average’, ’n clusters’: 7	0.50	0.52	0.19	0.09	0.00	7.00	False	False
’linkage’: ’average’, ’n clusters’: 8	0.78	0.78	0.27	0.10	0.00	8.00	True	False
’linkage’: ’average’, ’n clusters’: 9	0.77	0.78	0.26	0.11	0.00	9.00	False	False
’linkage’: ’average’, ’n clusters’: 10	0.76	0.77	0.21	0.11	0.00	10.00	False	False

Table C.2: Wine dataset - Hierarchical clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’linkage’: ’ward’, ’n clusters’: 2	0.63	0.63	0.47	0.14	0.06	2.00	True	True
’linkage’: ’ward’, ’n clusters’: 3	0.86	0.86	0.56	0.17	-0.07	3.00	True	True
’linkage’: ’ward’, ’n clusters’: 4	0.78	0.78	0.48	0.14	-0.03	4.00	False	False
’linkage’: ’ward’, ’n clusters’: 5	0.72	0.72	0.39	0.12	-0.21	5.00	False	False
’linkage’: ’ward’, ’n clusters’: 6	0.67	0.68	0.38	0.12	-0.26	6.00	False	False
’linkage’: ’ward’, ’n clusters’: 7	0.63	0.64	0.40	0.12	-0.12	7.00	False	False
’linkage’: ’ward’, ’n clusters’: 8	0.61	0.62	0.39	0.12	-0.19	8.00	False	False
’linkage’: ’ward’, ’n clusters’: 9	0.58	0.60	0.38	0.11	-0.14	9.00	False	False
’linkage’: ’ward’, ’n clusters’: 10	0.57	0.58	0.37	0.12	-0.10	10.00	False	False
’linkage’: ’single’, ’n clusters’: 2	0.00	0.01	0.19	0.14	0.00	2.00	False	False
’linkage’: ’single’, ’n clusters’: 3	0.00	0.02	0.02	0.16	0.00	3.00	False	False
’linkage’: ’single’, ’n clusters’: 4	0.01	0.04	-0.20	0.15	0.00	4.00	False	False
’linkage’: ’single’, ’n clusters’: 5	0.01	0.05	-0.30	0.11	0.00	5.00	False	False
’linkage’: ’single’, ’n clusters’: 6	0.01	0.06	-0.38	0.12	0.00	6.00	False	False
’linkage’: ’single’, ’n clusters’: 7	0.57	0.59	-0.12	0.15	0.00	7.00	False	False
’linkage’: ’single’, ’n clusters’: 8	0.59	0.61	-0.21	0.15	0.00	8.00	False	False
’linkage’: ’single’, ’n clusters’: 9	0.58	0.60	-0.20	0.15	0.00	9.00	False	False
’linkage’: ’single’, ’n clusters’: 10	0.57	0.60	-0.25	0.15	0.00	10.00	False	False
’linkage’: ’complete’, ’n clusters’: 2	0.49	0.49	0.47	0.12	-0.27	2.00	False	False
’linkage’: ’complete’, ’n clusters’: 3	0.70	0.70	0.51	0.16	-0.12	3.00	False	False
’linkage’: ’complete’, ’n clusters’: 4	0.67	0.68	0.46	0.14	-0.13	4.00	False	False
’linkage’: ’complete’, ’n clusters’: 5	0.63	0.63	0.36	0.12	-0.21	5.00	False	False
’linkage’: ’complete’, ’n clusters’: 6	0.68	0.69	0.37	0.12	-0.16	6.00	False	False
’linkage’: ’complete’, ’n clusters’: 7	0.64	0.65	0.39	0.12	-0.19	7.00	False	False
’linkage’: ’complete’, ’n clusters’: 8	0.63	0.64	0.38	0.13	0.00	8.00	False	False
’linkage’: ’complete’, ’n clusters’: 9	0.59	0.60	0.38	0.12	0.00	9.00	False	False
’linkage’: ’complete’, ’n clusters’: 10	0.57	0.58	0.35	0.11	0.00	10.00	False	False
’linkage’: ’average’, ’n clusters’: 2	0.63	0.63	0.47	0.14	0.06	2.00	False	False
’linkage’: ’average’, ’n clusters’: 3	0.78	0.78	0.55	0.17	-0.15	3.00	False	False
’linkage’: ’average’, ’n clusters’: 4	0.77	0.77	0.50	0.19	0.00	4.00	True	False
’linkage’: ’average’, ’n clusters’: 5	0.75	0.75	0.41	0.18	0.00	5.00	False	False
’linkage’: ’average’, ’n clusters’: 6	0.75	0.76	0.36	0.17	0.00	6.00	False	False
’linkage’: ’average’, ’n clusters’: 7	0.69	0.70	0.35	0.16	0.00	7.00	False	False
’linkage’: ’average’, ’n clusters’: 8	0.68	0.69	0.35	0.16	0.00	8.00	False	False
’linkage’: ’average’, ’n clusters’: 9	0.64	0.65	0.36	0.15	0.00	9.00	False	False
’linkage’: ’average’, ’n clusters’: 10	0.60	0.62	0.37	0.14	0.00	10.00	False	False

Table C.3: Wine dataset - Hierarchical clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’linkage’: ’ward’, ’n clusters’: 2	0.49	0.49	0.42	0.11	-0.25	2.00	False	False
’linkage’: ’ward’, ’n clusters’: 3	0.66	0.66	0.44	0.13	-0.33	3.00	False	False
’linkage’: ’ward’, ’n clusters’: 4	0.61	0.62	0.41	0.13	-0.39	4.00	False	False
’linkage’: ’ward’, ’n clusters’: 5	0.58	0.59	0.38	0.13	-0.30	5.00	False	False
’linkage’: ’ward’, ’n clusters’: 6	0.63	0.64	0.38	0.12	-0.21	6.00	False	False
’linkage’: ’ward’, ’n clusters’: 7	0.58	0.59	0.33	0.11	-0.32	7.00	False	False
’linkage’: ’ward’, ’n clusters’: 8	0.56	0.57	0.35	0.12	-0.30	8.00	False	False
’linkage’: ’ward’, ’n clusters’: 9	0.55	0.57	0.34	0.12	-0.23	9.00	False	False
’linkage’: ’ward’, ’n clusters’: 10	0.53	0.54	0.33	0.11	-0.26	10.00	False	False
’linkage’: ’single’, ’n clusters’: 2	-0.00	0.01	0.30	0.19	0.00	2.00	True	False
’linkage’: ’single’, ’n clusters’: 3	-0.00	0.02	0.09	0.19	0.00	3.00	False	False
’linkage’: ’single’, ’n clusters’: 4	-0.00	0.03	0.00	0.19	0.00	4.00	False	False
’linkage’: ’single’, ’n clusters’: 5	-0.01	0.03	-0.04	0.18	0.00	5.00	False	False
’linkage’: ’single’, ’n clusters’: 6	-0.01	0.04	-0.08	0.18	0.00	6.00	False	False
’linkage’: ’single’, ’n clusters’: 7	-0.00	0.06	-0.21	0.15	0.00	7.00	False	False
’linkage’: ’single’, ’n clusters’: 8	0.00	0.07	-0.20	0.15	0.00	8.00	False	False
’linkage’: ’single’, ’n clusters’: 9	0.00	0.07	-0.26	0.15	0.00	9.00	False	False
’linkage’: ’single’, ’n clusters’: 10	0.01	0.09	-0.24	0.15	0.00	10.00	False	False
’linkage’: ’complete’, ’n clusters’: 2	0.50	0.50	0.38	0.10	-0.48	2.00	False	False
’linkage’: ’complete’, ’n clusters’: 3	0.65	0.66	0.44	0.13	-0.27	3.00	True	True
’linkage’: ’complete’, ’n clusters’: 4	0.66	0.66	0.42	0.13	-0.12	4.00	True	True
’linkage’: ’complete’, ’n clusters’: 5	0.69	0.70	0.41	0.13	-0.15	5.00	True	True
’linkage’: ’complete’, ’n clusters’: 6	0.68	0.69	0.40	0.13	0.00	6.00	False	False
’linkage’: ’complete’, ’n clusters’: 7	0.69	0.69	0.35	0.14	0.00	7.00	False	False
’linkage’: ’complete’, ’n clusters’: 8	0.64	0.66	0.32	0.13	0.00	8.00	False	False
’linkage’: ’complete’, ’n clusters’: 9	0.60	0.61	0.31	0.13	0.00	9.00	False	False
’linkage’: ’complete’, ’n clusters’: 10	0.57	0.58	0.30	0.12	0.00	10.00	False	False
’linkage’: ’average’, ’n clusters’: 2	0.60	0.60	0.38	0.11	-0.34	2.00	False	False
’linkage’: ’average’, ’n clusters’: 3	0.78	0.78	0.46	0.13	-0.29	3.00	True	True
’linkage’: ’average’, ’n clusters’: 4	0.79	0.79	0.43	0.13	-0.27	4.00	True	True
’linkage’: ’average’, ’n clusters’: 5	0.73	0.73	0.41	0.13	-0.20	5.00	False	False
’linkage’: ’average’, ’n clusters’: 6	0.71	0.72	0.40	0.14	0.00	6.00	True	False
’linkage’: ’average’, ’n clusters’: 7	0.71	0.72	0.37	0.14	0.00	7.00	True	False
’linkage’: ’average’, ’n clusters’: 8	0.71	0.72	0.35	0.14	0.00	8.00	True	False
’linkage’: ’average’, ’n clusters’: 9	0.66	0.67	0.33	0.14	0.00	9.00	False	False
’linkage’: ’average’, ’n clusters’: 10	0.62	0.64	0.33	0.13	0.00	10.00	False	False

Table C.4: Wine dataset - Hierarchical clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’linkage’: ’ward’, ’n clusters’: 2	0.63	0.63	0.49	0.14	-0.03	2.00	False	False
’linkage’: ’ward’, ’n clusters’: 3	0.84	0.84	0.55	0.16	0.12	3.00	False	False
’linkage’: ’ward’, ’n clusters’: 4	0.76	0.76	0.49	0.14	-0.03	4.00	False	False
’linkage’: ’ward’, ’n clusters’: 5	0.71	0.71	0.49	0.14	-0.01	5.00	False	False
’linkage’: ’ward’, ’n clusters’: 6	0.68	0.68	0.44	0.14	-0.11	6.00	False	False
’linkage’: ’ward’, ’n clusters’: 7	0.63	0.64	0.38	0.12	-0.18	7.00	False	False
’linkage’: ’ward’, ’n clusters’: 8	0.61	0.62	0.40	0.12	-0.10	8.00	False	False
’linkage’: ’ward’, ’n clusters’: 9	0.58	0.59	0.39	0.12	-0.06	9.00	False	False
’linkage’: ’ward’, ’n clusters’: 10	0.56	0.58	0.37	0.12	-0.13	10.00	False	False
’linkage’: ’single’, ’n clusters’: 2	0.00	0.01	0.27	0.17	0.00	2.00	False	False
’linkage’: ’single’, ’n clusters’: 3	0.00	0.02	-0.05	0.14	0.00	3.00	False	False
’linkage’: ’single’, ’n clusters’: 4	0.00	0.03	-0.29	0.11	0.00	4.00	False	False
’linkage’: ’single’, ’n clusters’: 5	0.46	0.47	0.01	0.16	0.00	5.00	False	False
’linkage’: ’single’, ’n clusters’: 6	0.58	0.60	0.05	0.15	0.00	6.00	False	False
’linkage’: ’single’, ’n clusters’: 7	0.60	0.61	-0.03	0.16	0.00	7.00	False	False
’linkage’: ’single’, ’n clusters’: 8	0.59	0.60	-0.07	0.16	0.00	8.00	False	False
’linkage’: ’single’, ’n clusters’: 9	0.58	0.60	-0.12	0.16	0.00	9.00	False	False
’linkage’: ’single’, ’n clusters’: 10	0.56	0.58	-0.17	0.14	0.00	10.00	False	False
’linkage’: ’complete’, ’n clusters’: 2	0.59	0.59	0.49	0.14	-0.14	2.00	False	False
’linkage’: ’complete’, ’n clusters’: 3	0.83	0.84	0.55	0.16	0.24	3.00	True	True
’linkage’: ’complete’, ’n clusters’: 4	0.77	0.78	0.49	0.15	0.19	4.00	False	False
’linkage’: ’complete’, ’n clusters’: 5	0.71	0.71	0.48	0.13	0.05	5.00	False	False
’linkage’: ’complete’, ’n clusters’: 6	0.70	0.71	0.47	0.15	0.00	6.00	False	False
’linkage’: ’complete’, ’n clusters’: 7	0.64	0.65	0.34	0.11	0.00	7.00	False	False
’linkage’: ’complete’, ’n clusters’: 8	0.61	0.62	0.33	0.12	0.00	8.00	False	False
’linkage’: ’complete’, ’n clusters’: 9	0.59	0.60	0.34	0.13	0.00	9.00	False	False
’linkage’: ’complete’, ’n clusters’: 10	0.57	0.58	0.32	0.13	0.00	10.00	False	False
’linkage’: ’average’, ’n clusters’: 2	0.63	0.63	0.49	0.14	-0.03	2.00	False	False
’linkage’: ’average’, ’n clusters’: 3	0.86	0.86	0.55	0.16	0.19	3.00	True	True
’linkage’: ’average’, ’n clusters’: 4	0.85	0.85	0.51	0.19	0.00	4.00	True	False
’linkage’: ’average’, ’n clusters’: 5	0.77	0.77	0.47	0.16	0.00	5.00	False	False
’linkage’: ’average’, ’n clusters’: 6	0.71	0.72	0.49	0.16	0.00	6.00	False	False
’linkage’: ’average’, ’n clusters’: 7	0.67	0.68	0.45	0.16	0.00	7.00	False	False
’linkage’: ’average’, ’n clusters’: 8	0.64	0.65	0.41	0.14	0.00	8.00	False	False
’linkage’: ’average’, ’n clusters’: 9	0.62	0.63	0.42	0.15	0.00	9.00	False	False
’linkage’: ’average’, ’n clusters’: 10	0.62	0.63	0.40	0.15	0.00	10.00	False	False

Table C.5: Wine dataset - Hierarchical clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’linkage’: ’ward’, ’n clusters’: 2	0.61	0.61	0.53	0.14	0.36	2.00	False	False
’linkage’: ’ward’, ’n clusters’: 3	0.78	0.78	0.63	0.19	0.46	3.00	True	True
’linkage’: ’ward’, ’n clusters’: 4	0.70	0.70	0.56	0.14	0.23	4.00	False	False
’linkage’: ’ward’, ’n clusters’: 5	0.65	0.65	0.48	0.13	0.11	5.00	False	False
’linkage’: ’ward’, ’n clusters’: 6	0.61	0.62	0.41	0.11	-0.08	6.00	False	False
’linkage’: ’ward’, ’n clusters’: 7	0.59	0.60	0.42	0.13	-0.03	7.00	False	False
’linkage’: ’ward’, ’n clusters’: 8	0.56	0.58	0.45	0.14	0.02	8.00	False	False
’linkage’: ’ward’, ’n clusters’: 9	0.53	0.55	0.47	0.14	0.00	9.00	False	False
’linkage’: ’ward’, ’n clusters’: 10	0.51	0.53	0.45	0.12	0.05	10.00	False	False
’linkage’: ’single’, ’n clusters’: 2	0.61	0.61	0.53	0.14	0.36	2.00	False	False
’linkage’: ’single’, ’n clusters’: 3	0.78	0.78	0.63	0.19	0.46	3.00	False	False
’linkage’: ’single’, ’n clusters’: 4	0.79	0.80	0.56	0.21	0.48	4.00	True	True
’linkage’: ’single’, ’n clusters’: 5	0.76	0.76	0.45	0.18	0.40	5.00	False	False
’linkage’: ’single’, ’n clusters’: 6	0.74	0.74	0.32	0.17	0.24	6.00	False	False
’linkage’: ’single’, ’n clusters’: 7	0.68	0.69	0.36	0.16	0.30	7.00	False	False
’linkage’: ’single’, ’n clusters’: 8	0.67	0.68	0.30	0.16	0.00	8.00	False	False
’linkage’: ’single’, ’n clusters’: 9	0.66	0.67	0.31	0.16	0.00	9.00	False	False
’linkage’: ’single’, ’n clusters’: 10	0.61	0.62	0.30	0.14	0.00	10.00	False	False
’linkage’: ’complete’, ’n clusters’: 2	0.61	0.61	0.53	0.14	0.36	2.00	False	False
’linkage’: ’complete’, ’n clusters’: 3	0.78	0.78	0.63	0.19	0.46	3.00	False	False
’linkage’: ’complete’, ’n clusters’: 4	0.70	0.70	0.53	0.13	0.19	4.00	False	False
’linkage’: ’complete’, ’n clusters’: 5	0.66	0.66	0.44	0.10	-0.00	5.00	False	False
’linkage’: ’complete’, ’n clusters’: 6	0.62	0.63	0.47	0.12	0.05	6.00	False	False
’linkage’: ’complete’, ’n clusters’: 7	0.60	0.61	0.49	0.14	0.10	7.00	False	False
’linkage’: ’complete’, ’n clusters’: 8	0.56	0.57	0.43	0.13	-0.06	8.00	False	False
’linkage’: ’complete’, ’n clusters’: 9	0.58	0.60	0.44	0.14	0.03	9.00	False	False
’linkage’: ’complete’, ’n clusters’: 10	0.55	0.57	0.46	0.14	0.01	10.00	False	False
’linkage’: ’average’, ’n clusters’: 2	0.61	0.61	0.53	0.14	0.36	2.00	False	False
’linkage’: ’average’, ’n clusters’: 3	0.78	0.78	0.63	0.19	0.46	3.00	False	False
’linkage’: ’average’, ’n clusters’: 4	0.70	0.70	0.56	0.14	0.23	4.00	False	False
’linkage’: ’average’, ’n clusters’: 5	0.68	0.69	0.43	0.14	0.08	5.00	False	False
’linkage’: ’average’, ’n clusters’: 6	0.66	0.67	0.44	0.16	0.13	6.00	False	False
’linkage’: ’average’, ’n clusters’: 7	0.62	0.63	0.48	0.16	0.12	7.00	False	False
’linkage’: ’average’, ’n clusters’: 8	0.58	0.59	0.45	0.14	-0.00	8.00	False	False
’linkage’: ’average’, ’n clusters’: 9	0.60	0.61	0.45	0.15	0.09	9.00	False	False
’linkage’: ’average’, ’n clusters’: 10	0.57	0.58	0.46	0.15	0.10	10.00	False	False

Table C.6: Wine dataset - Hierarchical clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'linkage': 'ward', 'n clusters': 2	0.61	0.61	0.57	0.19	-0.20	2.00	False	False
'linkage': 'ward', 'n clusters': 3	0.80	0.80	0.65	0.22	-0.16	3.00	True	True
'linkage': 'ward', 'n clusters': 4	0.72	0.72	0.54	0.15	-0.15	4.00	False	False
'linkage': 'ward', 'n clusters': 5	0.67	0.68	0.44	0.11	-0.21	5.00	False	False
'linkage': 'ward', 'n clusters': 6	0.65	0.66	0.39	0.11	-0.14	6.00	False	False
'linkage': 'ward', 'n clusters': 7	0.61	0.62	0.34	0.10	-0.20	7.00	False	False
'linkage': 'ward', 'n clusters': 8	0.59	0.60	0.34	0.11	-0.19	8.00	False	False
'linkage': 'ward', 'n clusters': 9	0.59	0.60	0.34	0.12	-0.14	9.00	False	False
'linkage': 'ward', 'n clusters': 10	0.57	0.58	0.35	0.12	-0.16	10.00	False	False
'linkage': 'single', 'n clusters': 2	0.61	0.61	0.57	0.19	-0.20	2.00	False	False
'linkage': 'single', 'n clusters': 3	0.60	0.60	0.16	0.07	0.00	3.00	False	False
'linkage': 'single', 'n clusters': 4	0.60	0.61	-0.11	0.08	0.00	4.00	False	False
'linkage': 'single', 'n clusters': 5	0.58	0.59	-0.14	0.03	0.00	5.00	False	False
'linkage': 'single', 'n clusters': 6	0.60	0.61	-0.17	0.03	0.00	6.00	False	False
'linkage': 'single', 'n clusters': 7	0.77	0.78	0.22	0.17	0.00	7.00	False	False
'linkage': 'single', 'n clusters': 8	0.77	0.78	0.25	0.17	0.00	8.00	False	False
'linkage': 'single', 'n clusters': 9	0.76	0.77	0.20	0.17	0.00	9.00	False	False
'linkage': 'single', 'n clusters': 10	0.76	0.77	0.17	0.17	0.00	10.00	False	False
'linkage': 'complete', 'n clusters': 2	0.54	0.55	0.62	0.22	0.18	2.00	True	True
'linkage': 'complete', 'n clusters': 3	0.74	0.74	0.65	0.22	-0.04	3.00	True	True
'linkage': 'complete', 'n clusters': 4	0.76	0.76	0.54	0.19	-0.01	4.00	False	False
'linkage': 'complete', 'n clusters': 5	0.71	0.72	0.41	0.13	-0.04	5.00	False	False
'linkage': 'complete', 'n clusters': 6	0.65	0.66	0.39	0.11	-0.10	6.00	False	False
'linkage': 'complete', 'n clusters': 7	0.62	0.63	0.34	0.10	-0.16	7.00	False	False
'linkage': 'complete', 'n clusters': 8	0.58	0.59	0.36	0.11	-0.12	8.00	False	False
'linkage': 'complete', 'n clusters': 9	0.59	0.61	0.35	0.12	-0.10	9.00	False	False
'linkage': 'complete', 'n clusters': 10	0.57	0.59	0.36	0.12	-0.07	10.00	False	False
'linkage': 'average', 'n clusters': 2	0.54	0.55	0.62	0.22	0.18	2.00	False	False
'linkage': 'average', 'n clusters': 3	0.74	0.74	0.65	0.22	-0.04	3.00	False	False
'linkage': 'average', 'n clusters': 4	0.76	0.76	0.54	0.19	-0.01	4.00	False	False
'linkage': 'average', 'n clusters': 5	0.74	0.75	0.40	0.15	-0.04	5.00	False	False
'linkage': 'average', 'n clusters': 6	0.69	0.69	0.41	0.13	-0.02	6.00	False	False
'linkage': 'average', 'n clusters': 7	0.64	0.65	0.41	0.12	0.01	7.00	False	False
'linkage': 'average', 'n clusters': 8	0.60	0.61	0.36	0.11	-0.05	8.00	False	False
'linkage': 'average', 'n clusters': 9	0.58	0.59	0.37	0.12	-0.05	9.00	False	False
'linkage': 'average', 'n clusters': 10	0.55	0.57	0.36	0.12	-0.06	10.00	False	False

Table C.7: Wine dataset - k-means clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'n clusters': 2	0.50	0.48	0.26	0.07	-0.44	2.00	False	False
'n clusters': 3	0.87	0.88	0.28	0.07	-0.27	3.00	True	True
'n clusters': 4	0.81	0.79	0.26	0.06	-0.23	4.00	True	True
'n clusters': 5	0.72	0.72	0.23	0.06	-0.19	5.00	True	True
'n clusters': 6	0.65	0.61	0.19	0.06	-0.26	6.00	False	False
'n clusters': 7	0.62	0.61	0.19	0.05	-0.17	7.00	True	True
'n clusters': 8	0.64	0.63	0.18	0.06	-0.26	8.00	True	True
'n clusters': 9	0.59	0.56	0.16	0.06	-0.27	9.00	False	False
'n clusters': 10	0.59	0.59	0.14	0.06	-0.31	10.00	False	False
'n clusters': 11	0.54	0.58	0.14	0.06	-0.28	11.00	False	False

Table C.8: Wine dataset - k-means clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'n clusters': 2	0.47	0.48	0.46	0.11	-0.44	2.00	False	False
'n clusters': 3	0.88	0.87	0.56	0.17	-0.07	3.00	True	True
'n clusters': 4	0.72	0.73	0.49	0.13	-0.31	4.00	False	False
'n clusters': 5	0.69	0.67	0.45	0.13	-0.10	5.00	False	False
'n clusters': 6	0.61	0.63	0.44	0.13	-0.31	6.00	False	False
'n clusters': 7	0.62	0.63	0.42	0.13	-0.15	7.00	False	False
'n clusters': 8	0.60	0.60	0.41	0.12	-0.22	8.00	False	False
'n clusters': 9	0.58	0.62	0.41	0.12	-0.08	9.00	False	False
'n clusters': 10	0.58	0.58	0.40	0.12	-0.11	10.00	False	False
'n clusters': 11	0.54	0.57	0.40	0.12	-0.06	11.00	True	True

Table C.9: Wine dataset - *k*-means clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'n clusters': 2	0.46	0.47	0.41	0.10	-0.38	2.00	False	False
'n clusters': 3	0.83	0.83	0.48	0.14	-0.24	3.00	True	True
'n clusters': 4	0.71	0.69	0.43	0.13	-0.33	4.00	False	False
'n clusters': 5	0.65	0.69	0.41	0.12	-0.36	5.00	False	False
'n clusters': 6	0.61	0.64	0.34	0.11	-0.36	6.00	False	False
'n clusters': 7	0.57	0.61	0.37	0.12	-0.16	7.00	True	True
'n clusters': 8	0.55	0.56	0.37	0.12	-0.21	8.00	True	True
'n clusters': 9	0.56	0.58	0.35	0.12	-0.26	9.00	False	False
'n clusters': 10	0.54	0.55	0.34	0.11	-0.22	10.00	False	False
'n clusters': 11	0.53	0.54	0.34	0.11	-0.12	11.00	True	True

Table C.10: Wine dataset - *k*-means clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'n clusters': 2	0.49	0.50	0.50	0.13	0.08	2.00	False	False
'n clusters': 3	0.83	0.83	0.56	0.16	0.19	3.00	True	True
'n clusters': 4	0.78	0.78	0.51	0.14	-0.11	4.00	False	False
'n clusters': 5	0.71	0.70	0.48	0.14	-0.14	5.00	False	False
'n clusters': 6	0.65	0.65	0.47	0.14	-0.19	6.00	False	False
'n clusters': 7	0.63	0.63	0.44	0.13	-0.16	7.00	False	False
'n clusters': 8	0.61	0.64	0.44	0.14	-0.09	8.00	False	False
'n clusters': 9	0.59	0.61	0.42	0.12	-0.16	9.00	False	False
'n clusters': 10	0.57	0.61	0.42	0.12	-0.13	10.00	False	False
'n clusters': 11	0.58	0.56	0.41	0.13	-0.15	11.00	False	False

Table C.11: Wine dataset - *k*-means clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'n clusters': 2	0.51	0.60	0.50	0.13	-0.54	2.00	False	False
'n clusters': 3	0.78	0.78	0.63	0.19	0.46	3.00	True	True
'n clusters': 4	0.70	0.70	0.56	0.14	0.23	4.00	False	False
'n clusters': 5	0.66	0.66	0.47	0.12	0.06	5.00	False	False
'n clusters': 6	0.60	0.61	0.42	0.10	-0.09	6.00	False	False
'n clusters': 7	0.59	0.60	0.43	0.11	-0.01	7.00	False	False
'n clusters': 8	0.56	0.57	0.45	0.13	-0.12	8.00	False	False
'n clusters': 9	0.54	0.55	0.47	0.14	-0.05	9.00	False	False
'n clusters': 10	0.52	0.53	0.48	0.14	-0.02	10.00	False	False
'n clusters': 11	0.50	0.55	0.48	0.14	0.07	11.00	False	False

Table C.12: Wine dataset - *k*-means clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'n clusters': 2	0.47	0.47	0.60	0.19	-0.31	2.00	False	False
'n clusters': 3	0.79	0.80	0.66	0.22	0.03	3.00	True	True
'n clusters': 4	0.77	0.77	0.57	0.16	-0.06	4.00	False	False
'n clusters': 5	0.70	0.70	0.47	0.12	-0.11	5.00	False	False
'n clusters': 6	0.64	0.64	0.38	0.10	-0.09	6.00	False	False
'n clusters': 7	0.60	0.61	0.40	0.12	-0.08	7.00	False	False
'n clusters': 8	0.59	0.60	0.38	0.11	-0.02	8.00	False	False
'n clusters': 9	0.56	0.58	0.37	0.13	0.03	9.00	True	True
'n clusters': 10	0.55	0.56	0.38	0.12	-0.07	10.00	False	False
'n clusters': 11	0.53	0.54	0.39	0.13	-0.04	11.00	False	False

Table C.13: Wine dataset - *k*-medoids clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.45	0.45	0.26	0.07	-0.30	2.00	True	True
’n clusters’: 3	0.75	0.76	0.27	0.07	-0.37	3.00	True	True
’n clusters’: 4	0.69	0.69	0.20	0.05	-0.31	4.00	False	False
’n clusters’: 5	0.63	0.63	0.18	0.05	-0.36	5.00	False	False
’n clusters’: 6	0.62	0.63	0.17	0.06	-0.31	6.00	False	False
’n clusters’: 7	0.55	0.56	0.11	0.05	-0.35	7.00	False	False
’n clusters’: 8	0.55	0.56	0.14	0.05	-0.26	8.00	True	True
’n clusters’: 9	0.55	0.56	0.12	0.05	-0.27	9.00	True	True
’n clusters’: 10	0.51	0.53	0.12	0.05	-0.20	10.00	True	True
’n clusters’: 11	0.51	0.53	0.11	0.05	-0.20	11.00	True	True

Table C.14: Wine dataset - *k*-medoids clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.50	0.50	0.46	0.12	-0.31	2.00	False	False
’n clusters’: 3	0.84	0.85	0.56	0.17	-0.05	3.00	True	True
’n clusters’: 4	0.73	0.73	0.48	0.14	-0.21	4.00	False	False
’n clusters’: 5	0.68	0.69	0.42	0.12	-0.22	5.00	False	False
’n clusters’: 6	0.63	0.64	0.42	0.13	-0.03	6.00	True	True
’n clusters’: 7	0.65	0.65	0.36	0.10	-0.02	7.00	False	False
’n clusters’: 8	0.63	0.64	0.37	0.11	-0.01	8.00	False	False
’n clusters’: 9	0.61	0.62	0.40	0.13	-0.10	9.00	False	False
’n clusters’: 10	0.59	0.60	0.40	0.13	-0.02	10.00	True	True
’n clusters’: 11	0.57	0.59	0.39	0.13	0.00	11.00	True	False

Table C.15: Wine dataset - *k*-medoids clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.47	0.47	0.41	0.10	-0.10	2.00	True	True
’n clusters’: 3	0.74	0.75	0.47	0.14	-0.19	3.00	True	True
’n clusters’: 4	0.69	0.69	0.43	0.12	-0.33	4.00	False	False
’n clusters’: 5	0.63	0.64	0.37	0.11	-0.34	5.00	False	False
’n clusters’: 6	0.56	0.57	0.33	0.11	-0.46	6.00	False	False
’n clusters’: 7	0.54	0.56	0.30	0.10	-0.41	7.00	False	False
’n clusters’: 8	0.55	0.56	0.31	0.11	-0.30	8.00	False	False
’n clusters’: 9	0.54	0.55	0.30	0.11	-0.34	9.00	False	False
’n clusters’: 10	0.52	0.54	0.32	0.11	-0.15	10.00	True	True
’n clusters’: 11	0.53	0.54	0.32	0.12	-0.23	11.00	False	False

Table C.16: Wine dataset - *k*-medoids clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.49	0.50	0.48	0.12	-0.04	2.00	False	False
’n clusters’: 3	0.79	0.80	0.55	0.17	0.20	3.00	True	True
’n clusters’: 4	0.76	0.77	0.49	0.14	0.07	4.00	False	False
’n clusters’: 5	0.67	0.68	0.42	0.12	0.05	5.00	False	False
’n clusters’: 6	0.62	0.62	0.46	0.13	-0.00	6.00	False	False
’n clusters’: 7	0.63	0.64	0.42	0.12	-0.01	7.00	False	False
’n clusters’: 8	0.62	0.63	0.40	0.13	-0.00	8.00	False	False
’n clusters’: 9	0.59	0.61	0.35	0.12	-0.08	9.00	False	False
’n clusters’: 10	0.57	0.59	0.38	0.12	-0.15	10.00	False	False
’n clusters’: 11	0.55	0.57	0.39	0.13	-0.06	11.00	False	False

Table C.17: Wine dataset - *k*-medoids clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.61	0.61	0.53	0.14	0.36	2.00	False	False
’n clusters’: 3	0.78	0.78	0.63	0.19	0.46	3.00	True	True
’n clusters’: 4	0.70	0.70	0.55	0.14	0.18	4.00	False	False
’n clusters’: 5	0.66	0.66	0.48	0.11	0.01	5.00	False	False
’n clusters’: 6	0.69	0.70	0.46	0.12	-0.16	6.00	False	False
’n clusters’: 7	0.61	0.62	0.46	0.13	-0.18	7.00	False	False
’n clusters’: 8	0.59	0.61	0.46	0.12	-0.11	8.00	False	False
’n clusters’: 9	0.57	0.58	0.42	0.12	0.08	9.00	False	False
’n clusters’: 10	0.56	0.58	0.40	0.12	0.00	10.00	False	False
’n clusters’: 11	0.54	0.56	0.38	0.12	0.00	11.00	False	False

Table C.18: Wine dataset - *k*-medoids clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.47	0.47	0.60	0.19	-0.31	2.00	False	False
’n clusters’: 3	0.79	0.80	0.66	0.22	0.03	3.00	True	True
’n clusters’: 4	0.72	0.72	0.51	0.09	-0.10	4.00	False	False
’n clusters’: 5	0.71	0.72	0.52	0.13	0.06	5.00	True	True
’n clusters’: 6	0.70	0.71	0.50	0.14	-0.07	6.00	False	False
’n clusters’: 7	0.65	0.66	0.46	0.13	-0.08	7.00	False	False
’n clusters’: 8	0.61	0.62	0.40	0.12	-0.03	8.00	False	False
’n clusters’: 9	0.65	0.66	0.41	0.13	0.06	9.00	True	True
’n clusters’: 10	0.60	0.62	0.39	0.12	0.11	10.00	True	True
’n clusters’: 11	0.57	0.58	0.41	0.12	-0.05	11.00	False	False

Table C.19: Wine dataset - GMM clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’covariance type’: ’full’, ’n ’: 2	0.48	0.49	0.26	0.07	-0.45	2.00	False	False
’covariance type’: ’full’, ’n ’: 3	0.87	0.88	0.28	0.07	-0.27	3.00	True	True
’covariance type’: ’full’, ’n ’: 4	0.76	0.83	0.27	0.06	-0.14	4.00	True	True
’covariance type’: ’full’, ’n ’: 5	0.74	0.77	0.21	0.06	-0.33	5.00	False	False
’covariance type’: ’full’, ’n ’: 6	0.64	0.71	0.13	0.05	-0.28	6.00	False	False
’covariance type’: ’full’, ’n ’: 7	0.63	0.64	0.12	0.06	-0.32	7.00	False	False
’covariance type’: ’full’, ’n ’: 8	0.61	0.60	0.12	0.05	-0.35	8.00	False	False
’covariance type’: ’full’, ’n ’: 9	0.55	0.57	0.11	0.06	-0.31	9.00	False	False
’covariance type’: ’full’, ’n ’: 10	0.58	0.55	0.11	0.05	-0.25	10.00	False	False
’covariance type’: ’full’, ’n ’: 11	0.54	0.51	0.09	0.06	-0.25	11.00	False	False
’covariance type’: ’tied’, ’n ’: 2	0.49	0.49	0.27	0.06	-0.33	2.00	False	False
’covariance type’: ’tied’, ’n ’: 3	0.94	0.95	0.28	0.07	-0.33	3.00	True	True
’covariance type’: ’tied’, ’n ’: 4	0.83	0.78	0.24	0.06	-0.32	4.00	False	False
’covariance type’: ’tied’, ’n ’: 5	0.71	0.75	0.23	0.05	-0.23	5.00	False	False
’covariance type’: ’tied’, ’n ’: 6	0.76	0.84	0.19	0.06	-0.27	6.00	False	False
’covariance type’: ’tied’, ’n ’: 7	0.70	0.72	0.20	0.04	-0.30	7.00	False	False
’covariance type’: ’tied’, ’n ’: 8	0.71	0.67	0.10	0.05	-0.30	8.00	False	False
’covariance type’: ’tied’, ’n ’: 9	0.66	0.66	0.11	0.05	0.00	9.00	False	False
’covariance type’: ’tied’, ’n ’: 10	0.62	0.65	0.12	0.05	-0.27	10.00	False	False
’covariance type’: ’tied’, ’n ’: 11	0.62	0.60	0.11	0.05	0.00	11.00	False	False
’covariance type’: ’diag’, ’n ’: 2	0.46	0.51	0.26	0.07	-0.40	2.00	False	False
’covariance type’: ’diag’, ’n ’: 3	0.85	0.85	0.28	0.07	-0.29	3.00	False	False
’covariance type’: ’diag’, ’n ’: 4	0.78	0.75	0.25	0.06	-0.16	4.00	False	False
’covariance type’: ’diag’, ’n ’: 5	0.74	0.73	0.21	0.06	-0.19	5.00	False	False
’covariance type’: ’diag’, ’n ’: 6	0.67	0.65	0.18	0.05	-0.22	6.00	False	False
’covariance type’: ’diag’, ’n ’: 7	0.60	0.62	0.12	0.05	-0.21	7.00	False	False
’covariance type’: ’diag’, ’n ’: 8	0.61	0.65	0.11	0.06	-0.32	8.00	False	False
’covariance type’: ’diag’, ’n ’: 9	0.59	0.64	0.08	0.05	-0.26	9.00	False	False
’covariance type’: ’diag’, ’n ’: 10	0.59	0.66	0.10	0.05	-0.22	10.00	False	False
’covariance type’: ’diag’, ’n ’: 11	0.58	0.66	0.10	0.05	-0.26	11.00	False	False
’covariance type’: ’spherical’, ’n ’: 2	0.49	0.49	0.26	0.07	-0.32	2.00	True	True
’covariance type’: ’spherical’, ’n ’: 3	0.85	0.86	0.27	0.07	-0.27	3.00	True	True
’covariance type’: ’spherical’, ’n ’: 4	0.77	0.82	0.28	0.04	-0.26	4.00	True	True
’covariance type’: ’spherical’, ’n ’: 5	0.71	0.76	0.21	0.04	-0.22	5.00	False	False
’covariance type’: ’spherical’, ’n ’: 6	0.67	0.71	0.22	0.05	-0.24	6.00	False	False
’covariance type’: ’spherical’, ’n ’: 7	0.68	0.64	0.21	0.05	-0.17	7.00	False	False
’covariance type’: ’spherical’, ’n ’: 8	0.59	0.64	0.11	0.06	-0.26	8.00	True	True
’covariance type’: ’spherical’, ’n ’: 9	0.58	0.64	0.12	0.06	0.00	9.00	True	False
’covariance type’: ’spherical’, ’n ’: 10	0.55	0.61	0.13	0.06	-0.23	10.00	False	False
’covariance type’: ’spherical’, ’n ’: 11	0.57	0.61	0.11	0.06	-0.25	11.00	False	False

Table C.20: Wine dataset - GMM clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’covariance type’: ’full’, ’n ’: 2	0.49	0.49	0.47	0.12	-0.38	2.00	False	False
’covariance type’: ’full’, ’n ’: 3	0.84	0.88	0.56	0.16	-0.13	3.00	False	False
’covariance type’: ’full’, ’n ’: 4	0.69	0.69	0.49	0.14	-0.16	4.00	False	False
’covariance type’: ’full’, ’n ’: 5	0.68	0.69	0.46	0.12	-0.14	5.00	False	False
’covariance type’: ’full’, ’n ’: 6	0.68	0.67	0.41	0.12	-0.15	6.00	False	False
’covariance type’: ’full’, ’n ’: 7	0.65	0.65	0.35	0.12	-0.13	7.00	False	False
’covariance type’: ’full’, ’n ’: 8	0.58	0.65	0.39	0.13	-0.25	8.00	False	False
’covariance type’: ’full’, ’n ’: 9	0.61	0.62	0.37	0.11	0.09	9.00	True	True
’covariance type’: ’full’, ’n ’: 10	0.58	0.58	0.39	0.13	-0.05	10.00	False	False
’covariance type’: ’full’, ’n ’: 11	0.54	0.59	0.37	0.09	-0.19	11.00	False	False
’covariance type’: ’tied’, ’n ’: 2	0.64	0.48	0.46	0.12	0.06	2.00	True	True
’covariance type’: ’tied’, ’n ’: 3	0.88	0.88	0.56	0.17	0.02	3.00	True	True
’covariance type’: ’tied’, ’n ’: 4	0.77	0.77	0.50	0.14	-0.14	4.00	False	False
’covariance type’: ’tied’, ’n ’: 5	0.74	0.73	0.45	0.13	-0.15	5.00	False	False
’covariance type’: ’tied’, ’n ’: 6	0.62	0.69	0.41	0.13	-0.25	6.00	False	False
’covariance type’: ’tied’, ’n ’: 7	0.66	0.64	0.42	0.13	-0.11	7.00	False	False
’covariance type’: ’tied’, ’n ’: 8	0.60	0.62	0.39	0.13	-0.11	8.00	False	False
’covariance type’: ’tied’, ’n ’: 9	0.59	0.60	0.37	0.12	-0.11	9.00	False	False
’covariance type’: ’tied’, ’n ’: 10	0.56	0.57	0.40	0.12	-0.16	10.00	False	False
’covariance type’: ’tied’, ’n ’: 11	0.56	0.56	0.36	0.12	0.00	11.00	False	False
’covariance type’: ’diag’, ’n ’: 2	0.44	0.44	0.47	0.14	0.04	2.00	True	True
’covariance type’: ’diag’, ’n ’: 3	0.88	0.88	0.56	0.17	-0.07	3.00	True	True
’covariance type’: ’diag’, ’n ’: 4	0.71	0.72	0.49	0.15	-0.17	4.00	False	False
’covariance type’: ’diag’, ’n ’: 5	0.64	0.67	0.43	0.13	-0.46	5.00	False	False
’covariance type’: ’diag’, ’n ’: 6	0.59	0.68	0.41	0.13	-0.21	6.00	False	False
’covariance type’: ’diag’, ’n ’: 7	0.62	0.66	0.42	0.13	-0.19	7.00	False	False
’covariance type’: ’diag’, ’n ’: 8	0.61	0.63	0.39	0.13	-0.15	8.00	False	False
’covariance type’: ’diag’, ’n ’: 9	0.59	0.59	0.37	0.11	-0.17	9.00	False	False
’covariance type’: ’diag’, ’n ’: 10	0.58	0.59	0.35	0.11	-0.05	10.00	False	False
’covariance type’: ’diag’, ’n ’: 11	0.58	0.58	0.34	0.11	-0.02	11.00	False	False
’covariance type’: ’spherical’, ’n ’: 2	0.45	0.46	0.40	0.13	-0.47	2.00	False	False
’covariance type’: ’spherical’, ’n ’: 3	0.83	0.83	0.55	0.17	-0.09	3.00	False	False
’covariance type’: ’spherical’, ’n ’: 4	0.73	0.77	0.49	0.13	-0.17	4.00	False	False
’covariance type’: ’spherical’, ’n ’: 5	0.68	0.68	0.45	0.13	-0.03	5.00	False	False
’covariance type’: ’spherical’, ’n ’: 6	0.67	0.70	0.41	0.13	-0.19	6.00	False	False
’covariance type’: ’spherical’, ’n ’: 7	0.63	0.67	0.42	0.13	-0.07	7.00	False	False
’covariance type’: ’spherical’, ’n ’: 8	0.61	0.67	0.40	0.13	-0.15	8.00	False	False
’covariance type’: ’spherical’, ’n ’: 9	0.60	0.59	0.39	0.13	-0.10	9.00	False	False
’covariance type’: ’spherical’, ’n ’: 10	0.58	0.61	0.37	0.14	-0.12	10.00	False	False
’covariance type’: ’spherical’, ’n ’: 11	0.56	0.58	0.37	0.12	-0.10	11.00	False	False

Table C.21: Wine dataset - GMM clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’covariance type’: ’full’, ’n ’: 2	0.47	0.47	0.41	0.10	-0.37	2.00	False	False
’covariance type’: ’full’, ’n ’: 3	0.82	0.82	0.48	0.14	-0.28	3.00	False	False
’covariance type’: ’full’, ’n ’: 4	0.76	0.76	0.40	0.07	-0.28	4.00	False	False
’covariance type’: ’full’, ’n ’: 5	0.72	0.65	0.40	0.13	-0.23	5.00	False	False
’covariance type’: ’full’, ’n ’: 6	0.67	0.58	0.31	0.07	-0.28	6.00	False	False
’covariance type’: ’full’, ’n ’: 7	0.62	0.67	0.32	0.12	-0.37	7.00	False	False
’covariance type’: ’full’, ’n ’: 8	0.57	0.61	0.27	0.09	-0.32	8.00	False	False
’covariance type’: ’full’, ’n ’: 9	0.60	0.61	0.31	0.12	-0.28	9.00	False	False
’covariance type’: ’full’, ’n ’: 10	0.56	0.56	0.26	0.11	-0.27	10.00	False	False
’covariance type’: ’full’, ’n ’: 11	0.53	0.58	0.20	0.11	-0.09	11.00	False	False
’covariance type’: ’tied’, ’n ’: 2	0.47	0.47	0.41	0.10	-0.25	2.00	False	False
’covariance type’: ’tied’, ’n ’: 3	0.63	0.63	0.42	0.13	-0.30	3.00	False	False
’covariance type’: ’tied’, ’n ’: 4	0.77	0.66	0.43	0.13	-0.25	4.00	False	False
’covariance type’: ’tied’, ’n ’: 5	0.68	0.70	0.41	0.11	-0.27	5.00	False	False
’covariance type’: ’tied’, ’n ’: 6	0.65	0.71	0.36	0.13	-0.21	6.00	False	False
’covariance type’: ’tied’, ’n ’: 7	0.62	0.67	0.33	0.11	-0.30	7.00	False	False
’covariance type’: ’tied’, ’n ’: 8	0.65	0.63	0.34	0.11	-0.15	8.00	False	False
’covariance type’: ’tied’, ’n ’: 9	0.61	0.61	0.35	0.11	-0.15	9.00	True	True
’covariance type’: ’tied’, ’n ’: 10	0.59	0.60	0.29	0.12	-0.19	10.00	False	False
’covariance type’: ’tied’, ’n ’: 11	0.57	0.56	0.34	0.13	0.00	11.00	True	False
’covariance type’: ’diag’, ’n ’: 2	0.58	0.57	0.40	0.11	-0.45	2.00	False	False
’covariance type’: ’diag’, ’n ’: 3	0.80	0.81	0.48	0.14	-0.26	3.00	False	False
’covariance type’: ’diag’, ’n ’: 4	0.80	0.78	0.43	0.13	-0.17	4.00	True	True
’covariance type’: ’diag’, ’n ’: 5	0.69	0.74	0.39	0.13	-0.35	5.00	False	False
’covariance type’: ’diag’, ’n ’: 6	0.67	0.66	0.34	0.10	-0.22	6.00	False	False
’covariance type’: ’diag’, ’n ’: 7	0.72	0.66	0.36	0.11	-0.27	7.00	False	False
’covariance type’: ’diag’, ’n ’: 8	0.58	0.63	0.34	0.11	-0.23	8.00	False	False
’covariance type’: ’diag’, ’n ’: 9	0.62	0.59	0.32	0.11	0.00	9.00	False	False
’covariance type’: ’diag’, ’n ’: 10	0.58	0.56	0.29	0.11	-0.25	10.00	False	False
’covariance type’: ’diag’, ’n ’: 11	0.58	0.58	0.32	0.11	0.00	11.00	False	False
’covariance type’: ’spherical’, ’n ’: 2	0.48	0.48	0.41	0.10	-0.31	2.00	False	False
’covariance type’: ’spherical’, ’n ’: 3	0.82	0.85	0.48	0.14	-0.24	3.00	True	True
’covariance type’: ’spherical’, ’n ’: 4	0.72	0.76	0.43	0.13	-0.31	4.00	False	False
’covariance type’: ’spherical’, ’n ’: 5	0.68	0.73	0.40	0.12	-0.24	5.00	False	False
’covariance type’: ’spherical’, ’n ’: 6	0.73	0.66	0.38	0.13	-0.29	6.00	False	False
’covariance type’: ’spherical’, ’n ’: 7	0.62	0.68	0.34	0.12	-0.24	7.00	False	False
’covariance type’: ’spherical’, ’n ’: 8	0.60	0.61	0.35	0.12	-0.15	8.00	True	True
’covariance type’: ’spherical’, ’n ’: 9	0.56	0.59	0.31	0.11	-0.21	9.00	False	False
’covariance type’: ’spherical’, ’n ’: 10	0.55	0.57	0.33	0.11	-0.16	10.00	False	False
’covariance type’: ’spherical’, ’n ’: 11	0.53	0.53	0.31	0.12	0.00	11.00	False	False

Table C.22: Wine dataset - GMM clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’covariance type’: ’full’, ’n ’: 2	0.49	0.47	0.48	0.12	0.08	2.00	False	False
’covariance type’: ’full’, ’n ’: 3	0.86	0.87	0.55	0.16	0.48	3.00	True	True
’covariance type’: ’full’, ’n ’: 4	0.74	0.78	0.50	0.14	-0.04	4.00	False	False
’covariance type’: ’full’, ’n ’: 5	0.69	0.64	0.44	0.12	-0.04	5.00	False	False
’covariance type’: ’full’, ’n ’: 6	0.66	0.67	0.39	0.11	-0.15	6.00	False	False
’covariance type’: ’full’, ’n ’: 7	0.64	0.64	0.37	0.11	-0.12	7.00	False	False
’covariance type’: ’full’, ’n ’: 8	0.62	0.62	0.38	0.10	-0.10	8.00	False	False
’covariance type’: ’full’, ’n ’: 9	0.57	0.59	0.38	0.12	0.01	9.00	False	False
’covariance type’: ’full’, ’n ’: 10	0.58	0.60	0.29	0.12	-0.03	10.00	False	False
’covariance type’: ’full’, ’n ’: 11	0.56	0.57	0.25	0.11	-0.17	11.00	False	False
’covariance type’: ’tied’, ’n ’: 2	0.63	0.63	0.49	0.14	-0.14	2.00	False	False
’covariance type’: ’tied’, ’n ’: 3	0.83	0.83	0.56	0.16	0.19	3.00	True	True
’covariance type’: ’tied’, ’n ’: 4	0.79	0.71	0.49	0.14	-0.04	4.00	False	False
’covariance type’: ’tied’, ’n ’: 5	0.72	0.75	0.48	0.14	-0.20	5.00	False	False
’covariance type’: ’tied’, ’n ’: 6	0.69	0.71	0.48	0.14	-0.09	6.00	False	False
’covariance type’: ’tied’, ’n ’: 7	0.64	0.65	0.45	0.12	-0.19	6.00	False	False
’covariance type’: ’tied’, ’n ’: 8	0.62	0.64	0.40	0.12	0.00	8.00	False	False
’covariance type’: ’tied’, ’n ’: 9	0.60	0.60	0.41	0.14	-0.13	9.00	False	False
’covariance type’: ’tied’, ’n ’: 10	0.57	0.61	0.41	0.13	-0.08	10.00	False	False
’covariance type’: ’tied’, ’n ’: 11	0.59	0.59	0.36	0.10	0.00	11.00	False	False
’covariance type’: ’diag’, ’n ’: 2	0.56	0.56	0.49	0.13	-0.14	2.00	False	False
’covariance type’: ’diag’, ’n ’: 3	0.88	0.88	0.55	0.16	0.19	3.00	False	False
’covariance type’: ’diag’, ’n ’: 4	0.80	0.78	0.48	0.14	-0.07	4.00	False	False
’covariance type’: ’diag’, ’n ’: 5	0.71	0.66	0.43	0.14	-0.15	5.00	False	False
’covariance type’: ’diag’, ’n ’: 6	0.61	0.66	0.49	0.15	-0.13	6.00	False	False
’covariance type’: ’diag’, ’n ’: 7	0.64	0.65	0.46	0.12	-0.14	7.00	False	False
’covariance type’: ’diag’, ’n ’: 8	0.61	0.64	0.41	0.13	0.00	8.00	False	False
’covariance type’: ’diag’, ’n ’: 9	0.58	0.63	0.32	0.14	-0.15	9.00	False	False
’covariance type’: ’diag’, ’n ’: 10	0.57	0.56	0.37	0.10	-0.00	10.00	False	False
’covariance type’: ’diag’, ’n ’: 11	0.55	0.58	0.36	0.11	-0.09	11.00	False	False
’covariance type’: ’spherical’, ’n ’: 2	0.61	0.61	0.49	0.12	-0.03	2.00	False	False
’covariance type’: ’spherical’, ’n ’: 3	0.77	0.77	0.54	0.16	-0.11	3.00	False	False
’covariance type’: ’spherical’, ’n ’: 4	0.79	0.80	0.48	0.15	0.04	4.00	False	False
’covariance type’: ’spherical’, ’n ’: 5	0.72	0.75	0.48	0.15	-0.19	5.00	False	False
’covariance type’: ’spherical’, ’n ’: 6	0.68	0.70	0.42	0.15	-0.16	6.00	False	False
’covariance type’: ’spherical’, ’n ’: 7	0.62	0.66	0.44	0.12	-0.09	7.00	False	False
’covariance type’: ’spherical’, ’n ’: 8	0.63	0.63	0.40	0.14	-0.15	8.00	False	False
’covariance type’: ’spherical’, ’n ’: 9	0.60	0.60	0.37	0.12	-0.06	9.00	False	False
’covariance type’: ’spherical’, ’n ’: 10	0.60	0.61	0.37	0.15	-0.09	10.00	False	False
’covariance type’: ’spherical’, ’n ’: 11	0.56	0.57	0.35	0.13	0.02	11.00	False	False

Table C.23: Wine dataset - GMM clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’covariance type’: ’full’, ’n ’: 2	0.50	0.59	0.53	0.13	0.12	2.00	False	False
’covariance type’: ’full’, ’n ’: 3	0.78	0.78	0.63	0.19	0.46	3.00	True	True
’covariance type’: ’full’, ’n ’: 4	0.70	0.74	0.56	0.16	0.30	4.00	False	False
’covariance type’: ’full’, ’n ’: 5	0.67	0.65	0.48	0.10	0.07	5.00	False	False
’covariance type’: ’full’, ’n ’: 6	0.62	0.61	0.49	0.13	0.14	6.00	False	False
’covariance type’: ’full’, ’n ’: 7	0.58	0.59	0.44	0.10	0.17	7.00	False	False
’covariance type’: ’full’, ’n ’: 8	0.57	0.57	0.39	0.11	0.00	8.00	False	False
’covariance type’: ’full’, ’n ’: 9	0.56	0.56	0.41	0.12	-0.06	9.00	False	False
’covariance type’: ’full’, ’n ’: 10	0.52	0.57	0.42	0.12	-0.14	10.00	False	False
’covariance type’: ’full’, ’n ’: 11	0.51	0.51	0.46	0.12	-0.14	11.00	False	False
’covariance type’: ’tied’, ’n ’: 2	0.61	0.61	0.51	0.14	0.36	2.00	False	False
’covariance type’: ’tied’, ’n ’: 3	0.78	0.78	0.63	0.19	0.46	3.00	False	False
’covariance type’: ’tied’, ’n ’: 4	0.70	0.72	0.55	0.14	0.25	4.00	False	False
’covariance type’: ’tied’, ’n ’: 5	0.65	0.68	0.47	0.12	0.06	5.00	False	False
’covariance type’: ’tied’, ’n ’: 6	0.63	0.62	0.46	0.10	-0.09	6.00	False	False
’covariance type’: ’tied’, ’n ’: 7	0.58	0.62	0.40	0.09	-0.14	7.00	False	False
’covariance type’: ’tied’, ’n ’: 8	0.60	0.57	0.45	0.13	-0.06	8.00	False	False
’covariance type’: ’tied’, ’n ’: 9	0.54	0.56	0.47	0.14	-0.07	9.00	False	False
’covariance type’: ’tied’, ’n ’: 10	0.54	0.54	0.46	0.14	0.14	10.00	False	False
’covariance type’: ’tied’, ’n ’: 11	0.54	0.56	0.44	0.15	0.05	11.00	False	False
’covariance type’: ’diag’, ’n ’: 2	0.58	0.54	0.50	0.16	-0.09	2.00	False	False
’covariance type’: ’diag’, ’n ’: 3	0.78	0.78	0.63	0.19	0.46	3.00	False	False
’covariance type’: ’diag’, ’n ’: 4	0.70	0.71	0.53	0.14	0.22	4.00	False	False
’covariance type’: ’diag’, ’n ’: 5	0.65	0.67	0.46	0.11	0.09	5.00	False	False
’covariance type’: ’diag’, ’n ’: 6	0.64	0.66	0.45	0.13	-0.07	6.00	False	False
’covariance type’: ’diag’, ’n ’: 7	0.59	0.60	0.39	0.11	0.10	7.00	False	False
’covariance type’: ’diag’, ’n ’: 8	0.64	0.61	0.41	0.12	-0.02	8.00	False	False
’covariance type’: ’diag’, ’n ’: 9	0.59	0.57	0.47	0.13	0.10	9.00	False	False
’covariance type’: ’diag’, ’n ’: 10	0.54	0.59	0.45	0.13	0.02	10.00	False	False
’covariance type’: ’diag’, ’n ’: 11	0.52	0.54	0.45	0.14	-0.02	11.00	False	False
’covariance type’: ’spherical’, ’n ’: 2	0.65	0.65	0.49	0.15	-0.35	2.00	False	False
’covariance type’: ’spherical’, ’n ’: 3	0.78	0.78	0.63	0.19	0.46	3.00	False	False
’covariance type’: ’spherical’, ’n ’: 4	0.71	0.73	0.54	0.15	0.34	4.00	False	False
’covariance type’: ’spherical’, ’n ’: 5	0.67	0.67	0.41	0.14	0.05	5.00	False	False
’covariance type’: ’spherical’, ’n ’: 6	0.62	0.62	0.36	0.14	0.05	6.00	False	False
’covariance type’: ’spherical’, ’n ’: 7	0.59	0.61	0.42	0.12	-0.04	7.00	False	False
’covariance type’: ’spherical’, ’n ’: 8	0.57	0.58	0.39	0.13	-0.12	8.00	False	False
’covariance type’: ’spherical’, ’n ’: 9	0.54	0.55	0.48	0.14	-0.10	9.00	False	False
’covariance type’: ’spherical’, ’n ’: 10	0.53	0.59	0.48	0.14	0.03	10.00	False	False
’covariance type’: ’spherical’, ’n ’: 11	0.51	0.53	0.47	0.14	0.04	11.00	False	False

Table C.24: Wine dataset - GMM clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’covariance type’: ’full’, ’n ’: 2	0.64	0.65	0.59	0.22	-0.15	2.00	False	False
’covariance type’: ’full’, ’n ’: 3	0.86	0.86	0.65	0.22	-0.10	3.00	False	False
’covariance type’: ’full’, ’n ’: 4	0.75	0.76	0.57	0.16	-0.09	4.00	False	False
’covariance type’: ’full’, ’n ’: 5	0.66	0.69	0.39	0.13	-0.09	5.00	False	False
’covariance type’: ’full’, ’n ’: 6	0.64	0.64	0.37	0.12	-0.16	6.00	False	False
’covariance type’: ’full’, ’n ’: 7	0.62	0.61	0.42	0.11	-0.13	7.00	False	False
’covariance type’: ’full’, ’n ’: 8	0.57	0.59	0.32	0.12	-0.20	8.00	False	False
’covariance type’: ’full’, ’n ’: 9	0.58	0.59	0.32	0.12	0.07	9.00	False	False
’covariance type’: ’full’, ’n ’: 10	0.57	0.57	0.37	0.10	0.01	10.00	False	False
’covariance type’: ’full’, ’n ’: 11	0.53	0.55	0.34	0.10	-0.12	11.00	False	False
’covariance type’: ’tied’, ’n ’: 2	0.52	0.52	0.61	0.21	0.04	2.00	True	True
’covariance type’: ’tied’, ’n ’: 3	0.79	0.80	0.66	0.22	0.03	3.00	True	True
’covariance type’: ’tied’, ’n ’: 4	0.78	0.75	0.56	0.17	-0.02	4.00	False	False
’covariance type’: ’tied’, ’n ’: 5	0.79	0.73	0.43	0.11	-0.18	5.00	False	False
’covariance type’: ’tied’, ’n ’: 6	0.63	0.63	0.33	0.09	-0.09	6.00	False	False
’covariance type’: ’tied’, ’n ’: 7	0.64	0.66	0.29	0.11	-0.14	7.00	False	False
’covariance type’: ’tied’, ’n ’: 8	0.58	0.61	0.36	0.12	-0.06	8.00	False	False
’covariance type’: ’tied’, ’n ’: 9	0.58	0.60	0.35	0.12	-0.09	9.00	False	False
’covariance type’: ’tied’, ’n ’: 10	0.55	0.57	0.34	0.12	-0.03	10.00	False	False
’covariance type’: ’tied’, ’n ’: 11	0.56	0.55	0.37	0.12	-0.01	11.00	False	False
’covariance type’: ’diag’, ’n ’: 2	0.62	0.62	0.57	0.19	-0.20	2.00	False	False
’covariance type’: ’diag’, ’n ’: 3	0.83	0.83	0.65	0.22	-0.18	3.00	False	False
’covariance type’: ’diag’, ’n ’: 4	0.75	0.74	0.57	0.12	-0.13	4.00	False	False
’covariance type’: ’diag’, ’n ’: 5	0.70	0.68	0.45	0.12	-0.06	5.00	False	False
’covariance type’: ’diag’, ’n ’: 6	0.62	0.63	0.37	0.13	-0.13	6.00	False	False
’covariance type’: ’diag’, ’n ’: 7	0.62	0.61	0.38	0.12	-0.05	7.00	False	False
’covariance type’: ’diag’, ’n ’: 8	0.60	0.58	0.35	0.10	0.01	8.00	False	False
’covariance type’: ’diag’, ’n ’: 9	0.56	0.58	0.35	0.10	-0.08	9.00	False	False
’covariance type’: ’diag’, ’n ’: 10	0.55	0.55	0.37	0.12	-0.05	10.00	False	False
’covariance type’: ’diag’, ’n ’: 11	0.52	0.55	0.35	0.13	0.03	11.00	False	False
’covariance type’: ’spherical’, ’n ’: 2	0.60	0.60	0.60	0.22	-0.14	2.00	False	False
’covariance type’: ’spherical’, ’n ’: 3	0.79	0.80	0.66	0.22	0.03	3.00	False	False
’covariance type’: ’spherical’, ’n ’: 4	0.75	0.73	0.52	0.13	0.08	4.00	True	True
’covariance type’: ’spherical’, ’n ’: 5	0.70	0.73	0.42	0.12	-0.08	5.00	False	False
’covariance type’: ’spherical’, ’n ’: 6	0.68	0.66	0.37	0.13	-0.02	6.00	False	False
’covariance type’: ’spherical’, ’n ’: 7	0.61	0.62	0.36	0.14	-0.07	7.00	False	False
’covariance type’: ’spherical’, ’n ’: 8	0.60	0.59	0.38	0.13	-0.09	8.00	False	False
’covariance type’: ’spherical’, ’n ’: 9	0.59	0.59	0.37	0.13	-0.06	9.00	False	False
’covariance type’: ’spherical’, ’n ’: 10	0.53	0.58	0.34	0.12	0.02	10.00	False	False
’covariance type’: ’spherical’, ’n ’: 11	0.54	0.56	0.34	0.13	-0.05	11.00	False	False

Table C.25: Wine dataset - DBSCAN clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’eps’: 2.7, ’min samples’: 10	0.12	0.12	0.14	0.02	nan	2.00	False	False
’eps’: 2.7, ’min samples’: 20	0.48	0.49	0.20	0.04	0.08	3.00	False	False
’eps’: 2.7, ’min samples’: 30	0.50	0.50	0.18	0.06	nan	2.00	True	False
’eps’: 2.7, ’min samples’: 40	0.35	0.36	0.09	0.05	nan	2.00	False	False
’eps’: 2.7, ’min samples’: 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 2.7, ’min samples’: 60	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 2.8, ’min samples’: 10	0.10	0.10	0.16	0.02	nan	2.00	False	False
’eps’: 2.8, ’min samples’: 20	0.51	0.52	0.22	0.03	-0.13	3.00	False	False
’eps’: 2.8, ’min samples’: 30	0.49	0.49	0.16	0.05	0.22	3.00	True	True
’eps’: 2.8, ’min samples’: 40	0.52	0.52	0.17	0.06	nan	2.00	False	False
’eps’: 2.8, ’min samples’: 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 2.8, ’min samples’: 60	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 2.9, ’min samples’: 10	0.06	0.07	0.19	0.02	nan	2.00	False	False
’eps’: 2.9, ’min samples’: 20	0.05	0.06	0.16	0.02	nan	2.00	False	False
’eps’: 2.9, ’min samples’: 30	0.48	0.49	0.21	0.03	0.02	3.00	True	True
’eps’: 2.9, ’min samples’: 40	0.33	0.34	0.17	0.05	nan	2.00	False	False
’eps’: 2.9, ’min samples’: 50	0.42	0.42	0.14	0.06	nan	2.00	False	False
’eps’: 2.9, ’min samples’: 60	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 3.0, ’min samples’: 10	0.05	0.06	0.20	0.03	nan	2.00	False	False
’eps’: 3.0, ’min samples’: 20	0.06	0.07	0.18	0.02	nan	2.00	False	False
’eps’: 3.0, ’min samples’: 30	0.51	0.51	0.23	0.03	-0.13	3.00	False	False
’eps’: 3.0, ’min samples’: 40	0.39	0.40	0.20	0.05	nan	2.00	False	False
’eps’: 3.0, ’min samples’: 50	0.41	0.41	0.15	0.05	nan	2.00	False	False
’eps’: 3.0, ’min samples’: 60	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 3.1, ’min samples’: 10	0.04	0.05	0.23	0.03	nan	2.00	True	False
’eps’: 3.1, ’min samples’: 20	0.06	0.06	0.20	0.03	nan	2.00	False	False
’eps’: 3.1, ’min samples’: 30	0.48	0.49	0.23	0.03	-0.21	3.00	False	False
’eps’: 3.1, ’min samples’: 40	0.53	0.54	0.21	0.04	0.11	3.00	True	True
’eps’: 3.1, ’min samples’: 50	0.37	0.38	0.20	0.05	nan	2.00	False	False
’eps’: 3.1, ’min samples’: 60	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 3.2, ’min samples’: 10	0.04	0.05	0.23	0.03	nan	2.00	False	False
’eps’: 3.2, ’min samples’: 20	0.05	0.06	0.22	0.03	nan	2.00	False	False
’eps’: 3.2, ’min samples’: 30	0.05	0.06	0.22	0.03	nan	2.00	False	False
’eps’: 3.2, ’min samples’: 40	0.58	0.59	0.24	0.03	-0.13	3.00	True	True
’eps’: 3.2, ’min samples’: 50	0.41	0.41	0.22	0.05	nan	2.00	False	False
’eps’: 3.2, ’min samples’: 60	0.31	0.31	0.18	0.05	nan	2.00	False	False
’eps’: 3.3, ’min samples’: 10	0.04	0.05	0.23	0.03	nan	2.00	False	False
’eps’: 3.3, ’min samples’: 20	0.04	0.05	0.23	0.03	nan	2.00	False	False
’eps’: 3.3, ’min samples’: 30	0.05	0.06	0.22	0.03	nan	2.00	False	False
’eps’: 3.3, ’min samples’: 40	0.59	0.60	0.25	0.03	-0.14	3.00	True	True
’eps’: 3.3, ’min samples’: 50	0.44	0.44	0.23	0.05	nan	2.00	True	False
’eps’: 3.3, ’min samples’: 60	0.33	0.33	0.19	0.05	nan	2.00	False	False

Table C.26: Wine dataset - DBSCAN clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’eps’: 0.3, ’min samples’: 5	0.34	0.35	-0.11	0.10	0.28	6.00	True	True
’eps’: 0.3, ’min samples’: 10	0.07	0.08	-0.03	0.08	nan	2.00	False	False
’eps’: 0.3, ’min samples’: 15	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.3, ’min samples’: 20	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.3, ’min samples’: 25	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.3, ’min samples’: 30	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.3, ’min samples’: 35	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.3, ’min samples’: 40	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.3, ’min samples’: 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.3, ’min samples’: 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.5, ’min samples’: 5	0.49	0.50	0.32	0.07	-0.11	6.00	False	False
’eps’: 0.5, ’min samples’: 10	0.39	0.40	0.04	0.09	0.36	5.00	True	True
’eps’: 0.5, ’min samples’: 15	0.26	0.27	-0.05	0.09	0.19	3.00	False	False
’eps’: 0.5, ’min samples’: 20	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.9, ’min samples’: 5	-0.00	0.01	0.23	0.01	nan	2.00	False	False
’eps’: 0.9, ’min samples’: 10	-0.01	0.00	0.26	0.04	nan	2.00	False	False
’eps’: 0.9, ’min samples’: 15	0.60	0.61	0.39	0.02	0.22	3.00	True	True
’eps’: 0.9, ’min samples’: 20	0.59	0.59	0.36	0.07	0.34	4.00	True	True
’eps’: 0.9, ’min samples’: 25	0.47	0.47	0.27	0.09	0.31	3.00	True	True
’eps’: 0.9, ’min samples’: 30	0.34	0.34	0.32	0.12	nan	2.00	False	False
’eps’: 0.9, ’min samples’: 35	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.9, ’min samples’: 40	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.9, ’min samples’: 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.9, ’min samples’: 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1.1, ’min samples’: 5	0.00	0.01	0.19	0.14	nan	2.00	False	False
’eps’: 1.1, ’min samples’: 10	-0.00	0.01	0.23	0.01	nan	2.00	False	False
’eps’: 1.1, ’min samples’: 15	-0.00	0.01	0.23	0.01	nan	2.00	False	False
’eps’: 1.1, ’min samples’: 20	0.59	0.59	0.37	0.01	0.24	3.00	True	False
’eps’: 1.1, ’min samples’: 25	0.68	0.68	0.47	0.06	0.17	4.00	True	True
’eps’: 1.1, ’min samples’: 30	0.61	0.62	0.34	0.07	0.25	4.00	False	False
’eps’: 1.1, ’min samples’: 35	0.50	0.50	0.39	0.11	nan	2.00	True	False
’eps’: 1.1, ’min samples’: 40	0.32	0.33	0.31	0.12	nan	2.00	False	False
’eps’: 1.1, ’min samples’: 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1.1, ’min samples’: 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1.3, ’min samples’: 5	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.3, ’min samples’: 10	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.3, ’min samples’: 15	0.00	0.01	0.19	0.14	nan	2.00	False	False
’eps’: 1.3, ’min samples’: 20	0.00	0.01	0.19	0.14	nan	2.00	False	False
’eps’: 1.3, ’min samples’: 25	0.00	0.01	0.19	0.14	nan	2.00	False	False
’eps’: 1.3, ’min samples’: 30	0.63	0.63	0.37	0.04	0.04	3.00	False	False
’eps’: 1.3, ’min samples’: 35	0.76	0.76	0.48	0.06	0.14	4.00	True	True
’eps’: 1.3, ’min samples’: 40	0.54	0.54	0.39	0.11	nan	2.00	False	False
’eps’: 1.3, ’min samples’: 45	0.54	0.54	0.39	0.11	nan	2.00	False	False
’eps’: 1.3, ’min samples’: 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1.5, ’min samples’: 5	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.5, ’min samples’: 10	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.5, ’min samples’: 15	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.5, ’min samples’: 20	0.00	0.01	0.19	0.14	nan	2.00	False	False
’eps’: 1.5, ’min samples’: 25	0.00	0.01	0.19	0.14	nan	2.00	False	False
’eps’: 1.5, ’min samples’: 30	0.00	0.01	0.19	0.14	nan	2.00	False	False
’eps’: 1.5, ’min samples’: 35	0.59	0.59	0.37	0.16	0.04	3.00	True	True
’eps’: 1.5, ’min samples’: 40	0.73	0.74	0.48	0.07	-0.14	4.00	True	True
’eps’: 1.5, ’min samples’: 45	0.62	0.62	0.44	0.10	0.11	3.00	True	True
’eps’: 1.5, ’min samples’: 50	0.56	0.56	0.40	0.11	nan	2.00	False	False
’eps’: 1.7, ’min samples’: 5	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.7, ’min samples’: 10	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.7, ’min samples’: 15	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.7, ’min samples’: 20	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.7, ’min samples’: 25	0.00	0.01	0.19	0.14	nan	2.00	False	False
’eps’: 1.7, ’min samples’: 30	0.00	0.01	0.19	0.14	nan	2.00	False	False
’eps’: 1.7, ’min samples’: 35	0.00	0.01	0.19	0.14	nan	2.00	False	False
’eps’: 1.7, ’min samples’: 40	0.00	0.01	0.19	0.14	nan	2.00	False	False
’eps’: 1.7, ’min samples’: 45	0.60	0.61	0.35	0.06	0.04	3.00	False	False
’eps’: 1.7, ’min samples’: 50	0.58	0.58	0.45	0.11	nan	2.00	True	False

Table C.27: Wine dataset - DBSCAN clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'eps': 0.5, 'min samples': 5	0.29	0.32	-0.21	0.07	0.17	9.00	True	True
'eps': 0.5, 'min samples': 10	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.5, 'min samples': 15	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.5, 'min samples': 20	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.5, 'min samples': 25	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.5, 'min samples': 30	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.5, 'min samples': 35	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.5, 'min samples': 40	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.5, 'min samples': 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.5, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.7, 'min samples': 5	0.65	0.66	0.33	0.04	0.24	4.00	True	True
'eps': 0.7, 'min samples': 10	0.09	0.10	0.03	0.10	nan	2.00	False	False
'eps': 0.7, 'min samples': 15	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.7, 'min samples': 20	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.7, 'min samples': 25	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.7, 'min samples': 30	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.7, 'min samples': 35	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.7, 'min samples': 40	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.7, 'min samples': 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.7, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.9, 'min samples': 5	0.02	0.02	0.22	0.03	nan	2.00	False	False
'eps': 0.9, 'min samples': 10	0.60	0.61	0.27	0.06	0.41	4.00	True	True
'eps': 0.9, 'min samples': 15	0.22	0.22	0.09	0.09	nan	2.00	False	False
'eps': 0.9, 'min samples': 20	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.9, 'min samples': 25	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.9, 'min samples': 30	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.9, 'min samples': 35	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.9, 'min samples': 40	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.9, 'min samples': 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.9, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.1, 'min samples': 5	0.02	0.03	0.25	0.03	nan	2.00	False	False
'eps': 1.1, 'min samples': 10	0.51	0.52	0.29	0.01	-0.17	3.00	False	False
'eps': 1.1, 'min samples': 15	0.55	0.56	0.25	0.06	0.43	4.00	True	True
'eps': 1.1, 'min samples': 20	0.23	0.23	0.12	0.09	nan	2.00	False	False
'eps': 1.1, 'min samples': 25	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.1, 'min samples': 30	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.1, 'min samples': 35	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.1, 'min samples': 40	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.1, 'min samples': 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.1, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.3, 'min samples': 5	0.04	0.05	0.30	0.03	nan	2.00	False	False
'eps': 1.3, 'min samples': 10	0.03	0.04	0.27	0.03	nan	2.00	False	False
'eps': 1.3, 'min samples': 15	0.52	0.53	0.35	0.03	0.11	3.00	True	True
'eps': 1.3, 'min samples': 20	0.60	0.61	0.29	0.05	0.16	4.00	True	True
'eps': 1.3, 'min samples': 25	0.43	0.44	0.22	0.08	nan	2.00	False	False
'eps': 1.3, 'min samples': 30	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.3, 'min samples': 35	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.3, 'min samples': 40	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.3, 'min samples': 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.3, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.5, 'min samples': 5	0.03	0.04	0.30	0.02	nan	2.00	False	False
'eps': 1.5, 'min samples': 20	0.72	0.72	0.43	0.04	-0.03	4.00	True	True
'eps': 1.5, 'min samples': 25	0.64	0.65	0.34	0.05	0.12	4.00	True	True
'eps': 1.5, 'min samples': 30	0.46	0.46	0.23	0.08	nan	2.00	False	False
'eps': 1.7, 'min samples': 30	0.68	0.68	0.39	0.04	0.02	4.00	True	True
'eps': 1.9, 'min samples': 5	-0.00	0.01	0.30	0.19	nan	2.00	True	False
'eps': 1.9, 'min samples': 10	0.01	0.02	0.34	0.00	nan	2.00	False	False
'eps': 1.9, 'min samples': 15	0.02	0.03	0.31	0.03	nan	2.00	False	False
'eps': 1.9, 'min samples': 20	0.02	0.03	0.31	0.03	nan	2.00	False	False
'eps': 1.9, 'min samples': 25	0.02	0.03	0.27	0.02	nan	2.00	False	False
'eps': 1.9, 'min samples': 30	0.04	0.05	0.24	0.02	nan	2.00	False	False
'eps': 1.9, 'min samples': 35	0.69	0.70	0.41	0.04	-0.09	4.00	True	True
'eps': 2.1, 'min samples': 45	0.53	0.54	0.32	0.09	nan	2.00	True	False
'eps': 2.1, 'min samples': 50	0.52	0.52	0.30	0.09	nan	2.00	True	False

Table C.28: Wine dataset - DBSCAN clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'eps': 0.9, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.1, 'min samples': 5	0.00	0.01	0.27	0.17	nan	2.00	True	False
'eps': 1.1, 'min samples': 10	0.60	0.61	0.34	0.09	-0.12	3.00	False	False
'eps': 1.1, 'min samples': 15	0.53	0.54	0.25	0.07	0.14	4.00	False	False
'eps': 1.1, 'min samples': 20	0.53	0.54	0.31	0.08	0.48	4.00	True	True
'eps': 1.1, 'min samples': 25	0.45	0.46	0.38	0.13	nan	2.00	False	False
'eps': 1.1, 'min samples': 30	0.43	0.43	0.33	0.12	nan	2.00	False	False
'eps': 1.1, 'min samples': 35	0.34	0.34	0.27	0.12	nan	2.00	False	False
'eps': 1.1, 'min samples': 40	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.1, 'min samples': 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.1, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.3, 'min samples': 5	0.00	0.01	0.27	0.17	nan	2.00	False	False
'eps': 1.3, 'min samples': 10	0.00	0.01	0.27	0.17	nan	2.00	False	False
'eps': 1.3, 'min samples': 15	0.59	0.59	0.42	0.17	0.26	3.00	True	True
'eps': 1.3, 'min samples': 20	0.57	0.58	0.29	0.06	0.03	3.00	False	False
'eps': 1.3, 'min samples': 25	0.63	0.64	0.40	0.08	0.45	4.00	True	True
'eps': 1.3, 'min samples': 30	0.53	0.54	0.36	0.11	0.25	3.00	False	False
'eps': 1.3, 'min samples': 35	0.52	0.52	0.42	0.13	nan	2.00	False	False
'eps': 1.3, 'min samples': 40	0.46	0.46	0.35	0.12	nan	2.00	False	False
'eps': 1.3, 'min samples': 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.3, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.5, 'min samples': 5	0.00	0.01	0.27	0.17	nan	2.00	False	False
'eps': 1.5, 'min samples': 10	0.00	0.01	0.27	0.17	nan	2.00	False	False
'eps': 1.5, 'min samples': 15	0.00	0.01	0.27	0.17	nan	2.00	False	False
'eps': 1.5, 'min samples': 20	0.65	0.66	0.42	0.17	-0.03	3.00	True	True
'eps': 1.5, 'min samples': 25	0.76	0.76	0.48	0.08	0.31	4.00	True	True
'eps': 1.5, 'min samples': 30	0.62	0.63	0.48	0.15	0.22	3.00	True	True
'eps': 1.5, 'min samples': 35	0.66	0.67	0.44	0.12	nan	2.00	False	False
'eps': 1.5, 'min samples': 40	0.60	0.60	0.42	0.12	nan	2.00	False	False
'eps': 1.5, 'min samples': 45	0.50	0.50	0.38	0.12	nan	2.00	False	False
'eps': 1.5, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1.7, 'min samples': 5	0.00	0.01	0.27	0.17	nan	2.00	False	False
'eps': 1.7, 'min samples': 10	0.00	0.01	0.27	0.17	nan	2.00	False	False
'eps': 1.7, 'min samples': 15	0.00	0.01	0.27	0.17	nan	2.00	False	False
'eps': 1.7, 'min samples': 20	0.00	0.01	0.27	0.17	nan	2.00	False	False
'eps': 1.7, 'min samples': 25	0.61	0.61	0.40	0.15	-0.03	3.00	False	False
'eps': 1.7, 'min samples': 30	0.63	0.63	0.38	0.08	0.09	3.00	False	False
'eps': 1.7, 'min samples': 35	0.74	0.74	0.49	0.11	0.31	4.00	True	True
'eps': 1.7, 'min samples': 40	0.61	0.61	0.45	0.12	nan	2.00	False	False
'eps': 1.7, 'min samples': 45	0.62	0.62	0.44	0.12	nan	2.00	False	False
'eps': 1.7, 'min samples': 50	0.52	0.52	0.39	0.12	nan	2.00	False	False
'eps': 1.9, 'min samples': 5	-0.00	0.00	0.00	0.00	nan	1.00	False	False
'eps': 1.9, 'min samples': 10	-0.00	0.00	0.00	0.00	nan	1.00	False	False
'eps': 1.9, 'min samples': 15	0.00	0.01	0.27	0.17	nan	2.00	False	False
'eps': 1.9, 'min samples': 20	0.00	0.01	0.27	0.17	nan	2.00	False	False
'eps': 1.9, 'min samples': 25	0.00	0.01	0.27	0.17	nan	2.00	False	False
'eps': 1.9, 'min samples': 30	0.00	0.01	0.27	0.17	nan	2.00	False	False
'eps': 1.9, 'min samples': 35	0.58	0.59	0.39	0.15	0.26	3.00	False	False
'eps': 1.9, 'min samples': 40	0.56	0.56	0.50	0.14	nan	2.00	True	False
'eps': 1.9, 'min samples': 45	0.56	0.56	0.45	0.12	nan	2.00	False	False
'eps': 1.9, 'min samples': 50	0.58	0.59	0.45	0.12	nan	2.00	False	False
'eps': 2.1, 'min samples': 5	-0.00	0.00	0.00	0.00	nan	1.00	False	False
'eps': 2.1, 'min samples': 10	-0.00	0.00	0.00	0.00	nan	1.00	False	False
'eps': 2.1, 'min samples': 15	-0.00	0.00	0.00	0.00	nan	1.00	False	False
'eps': 2.1, 'min samples': 20	0.00	0.01	0.27	0.17	nan	2.00	False	False
'eps': 2.1, 'min samples': 25	0.00	0.01	0.27	0.17	nan	2.00	False	False
'eps': 2.1, 'min samples': 30	0.00	0.01	0.27	0.17	nan	2.00	False	False
'eps': 2.1, 'min samples': 35	0.00	0.01	0.27	0.17	nan	2.00	False	False
'eps': 2.1, 'min samples': 40	0.58	0.59	0.39	0.15	0.26	3.00	False	False
'eps': 2.1, 'min samples': 45	0.63	0.63	0.49	0.14	nan	2.00	True	False
'eps': 2.1, 'min samples': 50	0.61	0.61	0.49	0.14	nan	2.00	False	False

Table C.29: Wine dataset - DBSCAN clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’eps’: 10, ’min samples’: 5	0.79	0.80	0.56	0.21	0.54	4.00	True	True
’eps’: 10, ’min samples’: 10	0.75	0.75	0.56	0.04	0.53	4.00	False	False
’eps’: 12, ’min samples’: 15	0.75	0.75	0.57	0.08	0.58	4.00	True	True
’eps’: 14, ’min samples’: 10	0.79	0.80	0.56	0.21	0.54	4.00	False	False
’eps’: 14, ’min samples’: 15	0.80	0.80	0.57	0.15	0.56	4.00	True	True
’eps’: 14, ’min samples’: 20	0.80	0.80	0.57	0.15	0.56	4.00	False	False
’eps’: 14, ’min samples’: 25	0.77	0.77	0.55	0.10	0.58	4.00	True	True
’eps’: 14, ’min samples’: 30	0.62	0.62	0.44	0.08	0.53	4.00	False	False
’eps’: 14, ’min samples’: 35	0.29	0.29	0.06	0.04	nan	2.00	False	False
’eps’: 14, ’min samples’: 40	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 14, ’min samples’: 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 14, ’min samples’: 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 16, ’min samples’: 5	0.78	0.78	0.63	0.19	0.46	3.00	True	True
’eps’: 16, ’min samples’: 10	0.78	0.79	0.53	0.20	0.46	4.00	False	False
’eps’: 16, ’min samples’: 15	0.78	0.79	0.53	0.20	0.46	4.00	False	False
’eps’: 16, ’min samples’: 20	0.78	0.79	0.53	0.20	0.46	4.00	False	False
’eps’: 16, ’min samples’: 25	0.79	0.80	0.56	0.21	0.54	4.00	False	False
’eps’: 16, ’min samples’: 30	0.80	0.80	0.57	0.15	0.56	4.00	False	False
’eps’: 16, ’min samples’: 35	0.73	0.73	0.55	0.08	0.60	4.00	True	True
’eps’: 16, ’min samples’: 40	0.59	0.59	0.35	0.08	0.36	3.00	False	False
’eps’: 16, ’min samples’: 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 16, ’min samples’: 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 20, ’min samples’: 5	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 20, ’min samples’: 10	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 20, ’min samples’: 15	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 20, ’min samples’: 20	0.61	0.61	0.52	0.15	-0.22	2.00	False	False
’eps’: 20, ’min samples’: 25	0.61	0.61	0.52	0.15	-0.22	2.00	False	False
’eps’: 20, ’min samples’: 30	0.61	0.61	0.52	0.15	-0.22	2.00	False	False
’eps’: 20, ’min samples’: 35	0.81	0.81	0.62	0.19	-0.06	3.00	False	False
’eps’: 20, ’min samples’: 40	0.82	0.82	0.55	0.20	0.01	4.00	False	False
’eps’: 20, ’min samples’: 45	0.79	0.80	0.56	0.21	0.54	4.00	False	False
’eps’: 20, ’min samples’: 50	0.74	0.75	0.58	0.17	0.43	3.00	False	False
’eps’: 22, ’min samples’: 5	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 22, ’min samples’: 10	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 22, ’min samples’: 15	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 22, ’min samples’: 20	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 22, ’min samples’: 25	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 22, ’min samples’: 30	0.63	0.63	0.51	0.16	-0.35	2.00	False	False
’eps’: 22, ’min samples’: 35	0.63	0.63	0.51	0.16	-0.35	2.00	False	False
’eps’: 22, ’min samples’: 40	0.63	0.63	0.51	0.16	-0.35	2.00	False	False
’eps’: 22, ’min samples’: 45	0.83	0.83	0.55	0.20	-0.06	4.00	False	False
’eps’: 22, ’min samples’: 50	0.75	0.76	0.57	0.09	0.53	4.00	False	False

Table C.30: Wine dataset - DBSCAN clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’eps’: 2, ’min samples’: 5	-0.00	0.00	0.00	0.00	nan	1.00	True	False
’eps’: 2, ’min samples’: 10	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 2, ’min samples’: 15	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 2, ’min samples’: 20	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 2, ’min samples’: 25	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 2, ’min samples’: 30	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 2, ’min samples’: 35	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 2, ’min samples’: 40	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 2, ’min samples’: 45	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 2, ’min samples’: 50	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 3, ’min samples’: 5	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 3, ’min samples’: 10	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 3, ’min samples’: 15	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 3, ’min samples’: 20	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 3, ’min samples’: 25	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 3, ’min samples’: 30	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 3, ’min samples’: 35	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 3, ’min samples’: 40	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 3, ’min samples’: 45	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 3, ’min samples’: 50	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 4, ’min samples’: 5	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 4, ’min samples’: 10	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 4, ’min samples’: 15	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 4, ’min samples’: 20	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 4, ’min samples’: 25	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 4, ’min samples’: 30	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 4, ’min samples’: 35	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 4, ’min samples’: 40	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 4, ’min samples’: 45	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 4, ’min samples’: 50	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 5, ’min samples’: 5	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 5, ’min samples’: 10	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 5, ’min samples’: 15	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 5, ’min samples’: 20	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 5, ’min samples’: 25	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 5, ’min samples’: 30	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 5, ’min samples’: 35	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 5, ’min samples’: 40	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 5, ’min samples’: 45	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 5, ’min samples’: 50	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 6, ’min samples’: 5	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 6, ’min samples’: 10	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 6, ’min samples’: 15	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 6, ’min samples’: 20	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 6, ’min samples’: 25	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 6, ’min samples’: 30	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 6, ’min samples’: 35	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 6, ’min samples’: 40	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 6, ’min samples’: 45	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 6, ’min samples’: 50	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 7, ’min samples’: 5	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 7, ’min samples’: 10	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 7, ’min samples’: 15	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 7, ’min samples’: 20	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 7, ’min samples’: 25	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 7, ’min samples’: 30	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 7, ’min samples’: 35	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 7, ’min samples’: 40	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 7, ’min samples’: 45	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 7, ’min samples’: 50	-0.00	0.00	0.00	0.00	nan	1.00	False	False

Table C.31: Wine dataset - OPTICS clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'max eps': 2.7, 'min samples': 14	0.52	0.52	0.19	0.04	0.09	3.00	False	False
'max eps': 2.7, 'min samples': 15	0.52	0.52	0.19	0.04	0.09	3.00	True	True
'max eps': 2.7, 'min samples': 16	0.50	0.51	0.18	0.04	0.08	3.00	True	True
'max eps': 2.7, 'min samples': 17	0.50	0.51	0.18	0.04	0.08	3.00	False	False
'max eps': 2.7, 'min samples': 18	0.50	0.51	0.18	0.04	0.08	3.00	False	False
'max eps': 2.7, 'min samples': 19	0.43	0.44	0.15	0.03	0.08	3.00	False	False
'max eps': 2.7, 'min samples': 20	0.43	0.44	0.15	0.03	0.08	3.00	False	False
'max eps': 2.7, 'min samples': 21	0.43	0.44	0.15	0.03	0.08	3.00	False	False
'max eps': 2.8, 'min samples': 14	0.54	0.55	0.21	0.03	0.04	3.00	True	True
'max eps': 2.8, 'min samples': 15	0.54	0.55	0.21	0.03	0.04	3.00	False	False
'max eps': 2.8, 'min samples': 16	0.53	0.54	0.21	0.03	0.04	3.00	True	True
'max eps': 2.8, 'min samples': 17	0.53	0.54	0.21	0.03	0.04	3.00	False	False
'max eps': 2.8, 'min samples': 18	0.51	0.52	0.20	0.04	0.09	3.00	True	True
'max eps': 2.8, 'min samples': 19	0.51	0.52	0.20	0.04	0.09	3.00	True	True
'max eps': 2.8, 'min samples': 20	0.50	0.51	0.19	0.04	0.08	3.00	True	True
'max eps': 2.8, 'min samples': 21	0.50	0.51	0.19	0.04	0.08	3.00	False	False

Table C.32: Wine dataset - OPTICS clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'max eps': 1.2, 'min samples': 5	0.32	0.36	-0.09	0.07	0.32	11.00	False	False
'max eps': 1.2, 'min samples': 10	0.47	0.48	0.13	0.10	0.31	6.00	True	True
'max eps': 1.2, 'min samples': 15	0.37	0.38	0.02	0.09	0.32	4.00	True	True
'max eps': 1.2, 'min samples': 20	0.69	0.69	0.42	0.08	0.37	4.00	True	True
'max eps': 1.2, 'min samples': 25	0.67	0.67	0.38	0.07	0.50	4.00	False	False
'max eps': 1.2, 'min samples': 30	0.62	0.63	0.42	0.10	-0.01	3.00	False	False
'max eps': 1.2, 'min samples': 35	0.49	0.49	0.38	0.11	nan	2.00	False	False
'max eps': 1.2, 'min samples': 40	0.50	0.50	0.37	0.11	nan	2.00	False	False
'max eps': 1.2, 'min samples': 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.2, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.5, 'min samples': 5	0.32	0.36	-0.09	0.07	0.32	11.00	False	False
'max eps': 1.5, 'min samples': 10	0.47	0.48	0.13	0.10	0.31	6.00	False	False
'max eps': 1.5, 'min samples': 15	0.37	0.38	0.02	0.09	0.32	4.00	False	False
'max eps': 1.5, 'min samples': 20	0.69	0.69	0.42	0.08	0.37	4.00	False	False
'max eps': 1.5, 'min samples': 25	0.67	0.68	0.40	0.07	0.51	4.00	True	True
'max eps': 1.5, 'min samples': 30	0.67	0.67	0.39	0.06	0.37	4.00	False	False
'max eps': 1.5, 'min samples': 35	0.60	0.60	0.47	0.14	nan	2.00	True	False
'max eps': 1.5, 'min samples': 40	0.72	0.72	0.50	0.15	-0.19	3.00	True	True
'max eps': 1.5, 'min samples': 45	0.45	0.45	0.39	0.10	nan	2.00	False	False
'max eps': 1.5, 'min samples': 50	0.47	0.47	0.35	0.10	nan	2.00	False	False
'max eps': 1.8, 'min samples': 5	0.32	0.36	-0.09	0.07	0.32	11.00	False	False
'max eps': 1.8, 'min samples': 10	0.47	0.48	0.13	0.10	0.31	6.00	False	False
'max eps': 1.8, 'min samples': 15	0.37	0.38	0.02	0.09	0.32	4.00	False	False
'max eps': 1.8, 'min samples': 20	0.69	0.69	0.42	0.08	0.37	4.00	False	False
'max eps': 1.8, 'min samples': 25	0.70	0.70	0.40	0.07	0.51	4.00	True	True
'max eps': 1.8, 'min samples': 30	0.67	0.67	0.39	0.06	0.37	4.00	False	False
'max eps': 1.8, 'min samples': 35	0.00	0.01	0.19	0.14	nan	2.00	True	False
'max eps': 1.8, 'min samples': 40	0.50	0.50	0.41	0.11	nan	2.00	False	False

Table C.33: Wine dataset - OPTICS clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'max eps': 1.2, 'min samples': 5	0.25	0.27	-0.27	0.06	0.19	6.00	False	False
'max eps': 1.2, 'min samples': 10	0.34	0.35	-0.09	0.07	0.24	4.00	False	False
'max eps': 1.2, 'min samples': 15	0.66	0.66	0.31	0.05	0.37	4.00	True	True
'max eps': 1.2, 'min samples': 20	0.43	0.44	0.22	0.08	nan	2.00	False	False
'max eps': 1.2, 'min samples': 25	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.2, 'min samples': 30	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.2, 'min samples': 35	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.2, 'min samples': 40	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.2, 'min samples': 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.2, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.3, 'min samples': 5	0.25	0.27	-0.27	0.06	0.19	6.00	False	False
'max eps': 1.3, 'min samples': 10	0.34	0.35	-0.09	0.07	0.24	4.00	False	False
'max eps': 1.3, 'min samples': 15	0.58	0.58	0.29	0.08	0.21	3.00	True	True
'max eps': 1.3, 'min samples': 20	0.60	0.60	0.26	0.05	0.10	4.00	False	False
'max eps': 1.3, 'min samples': 25	0.42	0.43	0.21	0.08	nan	2.00	False	False
'max eps': 1.3, 'min samples': 30	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.3, 'min samples': 35	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.3, 'min samples': 40	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.3, 'min samples': 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.3, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.4, 'min samples': 5	0.25	0.27	-0.27	0.06	0.19	6.00	False	False
'max eps': 1.4, 'min samples': 10	0.34	0.35	-0.09	0.07	0.24	4.00	False	False
'max eps': 1.4, 'min samples': 15	0.56	0.57	0.27	0.08	0.24	3.00	True	True
'max eps': 1.4, 'min samples': 20	0.68	0.68	0.37	0.05	0.16	4.00	True	True
'max eps': 1.4, 'min samples': 25	0.60	0.61	0.20	0.08	0.29	3.00	True	True
'max eps': 1.4, 'min samples': 30	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.4, 'min samples': 35	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.4, 'min samples': 40	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.4, 'min samples': 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.4, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.5, 'min samples': 5	0.25	0.27	-0.27	0.06	0.19	6.00	False	False
'max eps': 1.5, 'min samples': 10	0.34	0.35	-0.09	0.07	0.24	4.00	False	False
'max eps': 1.5, 'min samples': 15	0.56	0.57	0.27	0.08	0.24	3.00	False	False
'max eps': 1.5, 'min samples': 20	0.71	0.72	0.42	0.04	0.03	4.00	True	True
'max eps': 1.5, 'min samples': 25	0.64	0.64	0.32	0.05	0.11	4.00	True	True
'max eps': 1.5, 'min samples': 30	0.45	0.45	0.22	0.08	nan	2.00	False	False
'max eps': 1.5, 'min samples': 35	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.5, 'min samples': 40	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.5, 'min samples': 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.5, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.6, 'min samples': 5	0.25	0.27	-0.27	0.06	0.19	6.00	False	False
'max eps': 1.6, 'min samples': 10	0.34	0.35	-0.09	0.07	0.24	4.00	False	False
'max eps': 1.6, 'min samples': 15	0.56	0.57	0.27	0.08	0.24	3.00	False	False
'max eps': 1.6, 'min samples': 20	0.58	0.59	0.33	0.09	nan	2.00	True	False
'max eps': 1.6, 'min samples': 25	0.69	0.69	0.38	0.04	0.06	4.00	False	False
'max eps': 1.9, 'min samples': 35	0.66	0.67	0.38	0.05	0.18	4.00	True	True
'max eps': 1.9, 'min samples': 40	0.48	0.49	0.27	0.08	nan	2.00	False	False
'max eps': 1.9, 'min samples': 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.9, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 2, 'min samples': 5	0.25	0.27	-0.27	0.06	0.19	6.00	False	False
'max eps': 2, 'min samples': 10	0.34	0.35	-0.09	0.07	0.24	4.00	False	False
'max eps': 2, 'min samples': 15	0.56	0.57	0.27	0.08	0.24	3.00	False	False
'max eps': 2, 'min samples': 20	0.58	0.59	0.33	0.09	nan	2.00	False	False
'max eps': 2, 'min samples': 25	0.03	0.04	0.28	0.02	nan	2.00	False	False
'max eps': 2, 'min samples': 30	0.02	0.02	0.25	0.03	nan	2.00	False	False
'max eps': 2, 'min samples': 35	0.02	0.02	0.21	0.02	nan	2.00	False	False
'max eps': 2, 'min samples': 40	0.49	0.49	0.29	0.08	nan	2.00	False	False
'max eps': 2, 'min samples': 45	0.38	0.38	0.20	0.07	nan	2.00	False	False
'max eps': 2, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False

Table C.34: Wine dataset - OPTICS clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'max eps': 1.2, 'min samples': 5	0.33	0.36	-0.03	0.08	0.23	12.00	False	False
'max eps': 1.2, 'min samples': 10	0.33	0.35	-0.10	0.08	0.25	5.00	False	False
'max eps': 1.2, 'min samples': 15	0.63	0.64	0.34	0.09	0.47	4.00	True	True
'max eps': 1.2, 'min samples': 20	0.59	0.60	0.33	0.09	0.40	4.00	False	False
'max eps': 1.2, 'min samples': 25	0.50	0.50	0.41	0.13	nan	2.00	False	False
'max eps': 1.2, 'min samples': 30	0.50	0.50	0.41	0.13	nan	2.00	False	False
'max eps': 1.2, 'min samples': 35	0.42	0.42	0.36	0.12	nan	2.00	False	False
'max eps': 1.2, 'min samples': 40	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.3, 'min samples': 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.3, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.5, 'min samples': 5	0.33	0.36	-0.03	0.08	0.23	12.00	False	False
'max eps': 1.5, 'min samples': 10	0.33	0.35	-0.10	0.08	0.25	5.00	False	False
'max eps': 1.5, 'min samples': 15	0.62	0.62	0.34	0.08	0.49	4.00	False	False
'max eps': 1.5, 'min samples': 20	0.68	0.68	0.41	0.07	0.51	4.00	True	True
'max eps': 1.5, 'min samples': 25	0.72	0.73	0.42	0.06	0.31	4.00	False	False
'max eps': 2, 'min samples': 10	0.33	0.35	-0.10	0.08	0.25	5.00	False	False
'max eps': 2, 'min samples': 15	0.63	0.64	0.34	0.08	0.49	4.00	True	True

Table C.35: Wine dataset - OPTICS clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'max eps': 15, 'min samples': 5	0.34	0.38	0.17	0.10	0.40	18.00	False	False
'max eps': 15, 'min samples': 10	0.42	0.43	0.09	0.07	0.21	8.00	False	False
'max eps': 15, 'min samples': 15	0.37	0.38	0.08	0.10	0.36	4.00	False	False
'max eps': 15, 'min samples': 20	0.58	0.59	0.31	0.07	0.43	4.00	False	False
'max eps': 15, 'min samples': 25	0.75	0.75	0.52	0.09	0.56	4.00	True	True
'max eps': 15, 'min samples': 30	0.63	0.64	0.41	0.08	0.52	4.00	False	False
'max eps': 15, 'min samples': 35	0.43	0.43	0.21	0.06	nan	2.00	False	False
'max eps': 15, 'min samples': 40	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 15, 'min samples': 45	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 15, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 17, 'min samples': 5	0.34	0.38	0.17	0.10	0.40	18.00	False	False
'max eps': 17, 'min samples': 10	0.42	0.43	0.09	0.07	0.21	8.00	False	False
'max eps': 17, 'min samples': 15	0.37	0.38	0.08	0.10	0.36	4.00	False	False
'max eps': 17, 'min samples': 20	0.59	0.59	0.31	0.07	0.44	4.00	False	False
'max eps': 17, 'min samples': 25	0.76	0.76	0.52	0.10	0.51	4.00	True	True
'max eps': 17, 'min samples': 30	0.76	0.77	0.52	0.09	0.56	4.00	True	True
'max eps': 17, 'min samples': 35	0.64	0.65	0.49	0.13	0.39	3.00	True	True
'max eps': 17, 'min samples': 40	0.62	0.62	0.47	0.12	0.39	3.00	False	False
'max eps': 17, 'min samples': 45	0.41	0.42	0.20	0.06	nan	2.00	False	False
'max eps': 17, 'min samples': 50	-0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 19, 'min samples': 5	0.34	0.38	0.17	0.10	0.40	18.00	False	False
'max eps': 19, 'min samples': 10	0.42	0.43	0.09	0.07	0.21	8.00	False	False
'max eps': 19, 'min samples': 15	0.37	0.38	0.08	0.10	0.36	4.00	False	False
'max eps': 19, 'min samples': 20	0.59	0.59	0.31	0.07	0.44	4.00	False	False
'max eps': 19, 'min samples': 25	0.81	0.81	0.44	0.18	-0.06	4.00	True	True
'max eps': 19, 'min samples': 30	0.76	0.77	0.50	0.09	0.51	4.00	False	False
'max eps': 19, 'min samples': 35	0.79	0.79	0.51	0.09	0.56	4.00	False	False
'max eps': 19, 'min samples': 40	0.75	0.76	0.50	0.08	0.55	4.00	False	False
'max eps': 19, 'min samples': 45	0.61	0.61	0.46	0.11	0.38	3.00	False	False
'max eps': 19, 'min samples': 50	0.56	0.56	0.40	0.10	0.38	3.00	False	False
'max eps': 21, 'min samples': 5	0.35	0.39	0.21	0.09	0.43	19.00	False	False
'max eps': 21, 'min samples': 10	0.44	0.45	0.19	0.06	0.24	8.00	False	False
'max eps': 21, 'min samples': 15	0.37	0.38	0.08	0.10	0.36	4.00	False	False
'max eps': 21, 'min samples': 20	0.58	0.59	0.31	0.07	0.45	4.00	False	False
'max eps': 21, 'min samples': 25	0.82	0.82	0.62	0.19	-0.13	3.00	True	True
'max eps': 21, 'min samples': 30	0.78	0.78	0.52	0.09	0.30	4.00	False	False
'max eps': 21, 'min samples': 35	0.77	0.77	0.49	0.10	0.32	4.00	False	False
'max eps': 21, 'min samples': 40	0.75	0.75	0.51	0.09	-0.14	4.00	False	False
'max eps': 21, 'min samples': 45	0.75	0.76	0.50	0.08	0.55	4.00	False	False
'max eps': 21, 'min samples': 50	0.63	0.64	0.48	0.12	0.39	3.00	False	False
'max eps': 23, 'min samples': 5	0.35	0.39	0.21	0.09	0.43	19.00	False	False
'max eps': 23, 'min samples': 10	0.44	0.45	0.19	0.06	0.24	8.00	False	False
'max eps': 23, 'min samples': 15	0.37	0.38	0.08	0.10	0.36	4.00	False	False
'max eps': 23, 'min samples': 20	0.58	0.59	0.31	0.07	0.45	4.00	False	False
'max eps': 23, 'min samples': 25	0.82	0.82	0.62	0.19	-0.13	3.00	False	False
'max eps': 23, 'min samples': 30	0.79	0.79	0.52	0.09	0.30	4.00	True	True
'max eps': 23, 'min samples': 35	0.77	0.77	0.50	0.09	0.31	4.00	False	False
'max eps': 23, 'min samples': 40	0.79	0.79	0.51	0.09	0.61	4.00	True	True
'max eps': 23, 'min samples': 45	0.76	0.76	0.36	0.14	-0.54	4.00	False	False
'max eps': 23, 'min samples': 50	0.68	0.68	0.53	0.16	-0.32	3.00	False	False
'max eps': 25, 'min samples': 5	0.35	0.39	0.21	0.09	0.43	19.00	False	False
'max eps': 25, 'min samples': 10	0.44	0.45	0.19	0.06	0.24	8.00	False	False
'max eps': 25, 'min samples': 15	0.37	0.38	0.08	0.10	0.36	4.00	False	False
'max eps': 25, 'min samples': 20	0.58	0.59	0.31	0.07	0.45	4.00	False	False
'max eps': 25, 'min samples': 25	0.82	0.82	0.62	0.19	-0.13	3.00	False	False
'max eps': 25, 'min samples': 30	0.79	0.79	0.52	0.09	0.30	4.00	False	False
'max eps': 25, 'min samples': 35	0.73	0.73	0.50	0.09	0.30	4.00	False	False
'max eps': 25, 'min samples': 40	0.75	0.75	0.51	0.09	0.64	4.00	True	True
'max eps': 25, 'min samples': 45	0.72	0.72	0.49	0.08	0.23	4.00	False	False
'max eps': 25, 'min samples': 50	0.70	0.70	0.56	0.17	-0.45	3.00	False	False

Table C.36: Wine dataset - OPTICS clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’max eps’: 4, ’min samples’: 5	0.27	0.32	-0.25	0.04	0.18	13.00	False	False
’max eps’: 4, ’min samples’: 10	0.30	0.31	-0.24	0.02	0.25	4.00	False	False
’max eps’: 4, ’min samples’: 15	0.74	0.74	0.52	0.13	0.05	4.00	False	False
’max eps’: 4, ’min samples’: 20	0.80	0.80	0.57	0.01	0.11	4.00	True	True
’max eps’: 4, ’min samples’: 25	0.80	0.80	0.58	0.05	0.06	4.00	True	True
’max eps’: 4, ’min samples’: 30	0.77	0.78	0.50	0.08	-0.03	4.00	False	False
’max eps’: 4, ’min samples’: 35	0.77	0.77	0.57	0.02	0.13	4.00	True	True
’max eps’: 4, ’min samples’: 40	0.73	0.73	0.55	0.04	0.33	4.00	True	True
’max eps’: 4, ’min samples’: 45	0.77	0.77	0.58	0.02	0.06	4.00	True	True
’max eps’: 4, ’min samples’: 50	0.70	0.70	0.55	0.15	0.29	3.00	True	True
’max eps’: 5, ’min samples’: 5	0.27	0.32	-0.25	0.04	0.18	13.00	False	False
’max eps’: 5, ’min samples’: 10	0.30	0.31	-0.24	0.02	0.25	4.00	False	False
’max eps’: 5, ’min samples’: 15	0.74	0.74	0.52	0.13	0.05	4.00	False	False
’max eps’: 5, ’min samples’: 20	0.80	0.80	0.57	0.01	0.11	4.00	False	False
’max eps’: 5, ’min samples’: 25	0.80	0.80	0.58	0.05	0.06	4.00	False	False

Table C.37: Wine dataset - HDBSCAN clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’min cluster size’: 5, ’min samples’: 5	0.17	0.19	-0.13	0.06	0.06	5.00	True	True
’min cluster size’: 5, ’min samples’: 10	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’min cluster size’: 5, ’min samples’: 15	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’min cluster size’: 5, ’min samples’: 20	0.25	0.26	0.15	0.07	nan	2.00	True	False
’min cluster size’: 5, ’min samples’: 25	0.45	0.46	0.22	0.07	nan	2.00	True	False
’min cluster size’: 5, ’min samples’: 30	0.33	0.33	0.17	0.07	nan	2.00	False	False
’min cluster size’: 5, ’min samples’: 35	0.33	0.33	0.18	0.07	nan	2.00	False	False
’min cluster size’: 10, ’min samples’: 5	0.52	0.53	0.22	0.07	nan	2.00	True	False
’min cluster size’: 10, ’min samples’: 10	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’min cluster size’: 10, ’min samples’: 15	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’min cluster size’: 10, ’min samples’: 20	0.25	0.26	0.15	0.07	nan	2.00	False	False
’min cluster size’: 10, ’min samples’: 25	0.45	0.46	0.22	0.07	nan	2.00	False	False
’min cluster size’: 10, ’min samples’: 30	0.33	0.33	0.17	0.07	nan	2.00	False	False
’min cluster size’: 10, ’min samples’: 35	0.33	0.33	0.18	0.07	nan	2.00	False	False
’min cluster size’: 15, ’min samples’: 5	0.52	0.53	0.22	0.07	nan	2.00	False	False
’min cluster size’: 15, ’min samples’: 10	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’min cluster size’: 15, ’min samples’: 15	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’min cluster size’: 15, ’min samples’: 20	0.25	0.26	0.15	0.07	nan	2.00	False	False
’min cluster size’: 15, ’min samples’: 25	0.45	0.46	0.22	0.07	nan	2.00	False	False
’min cluster size’: 15, ’min samples’: 30	0.33	0.33	0.17	0.07	nan	2.00	False	False
’min cluster size’: 15, ’min samples’: 35	0.33	0.33	0.18	0.07	nan	2.00	False	False

Table C.38: Wine dataset - HDBSCAN clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’min cluster size’: 10, ’min samples’: 5	0.42	0.44	0.06	0.10	0.29	6.00	False	False
’min cluster size’: 10, ’min samples’: 10	0.47	0.48	0.13	0.10	0.31	6.00	False	False
’min cluster size’: 10, ’min samples’: 15	0.43	0.44	0.05	0.10	0.30	5.00	False	False
’min cluster size’: 10, ’min samples’: 20	0.69	0.69	0.42	0.08	0.37	4.00	False	False
’min cluster size’: 10, ’min samples’: 25	0.70	0.70	0.40	0.07	0.51	4.00	True	True
’min cluster size’: 10, ’min samples’: 30	0.68	0.68	0.39	0.06	0.37	4.00	False	False
’min cluster size’: 10, ’min samples’: 35	0.41	0.41	0.37	0.13	nan	2.00	False	False
’min cluster size’: 20, ’min samples’: 5	0.45	0.45	0.24	0.07	-0.09	4.00	False	False
’min cluster size’: 20, ’min samples’: 10	0.53	0.54	0.25	0.07	0.36	4.00	False	False
’min cluster size’: 20, ’min samples’: 15	0.37	0.38	0.02	0.09	0.32	4.00	False	False
’min cluster size’: 20, ’min samples’: 20	0.69	0.69	0.42	0.08	0.37	4.00	False	False
’min cluster size’: 20, ’min samples’: 25	0.70	0.70	0.40	0.07	0.51	4.00	False	False
’min cluster size’: 20, ’min samples’: 30	0.68	0.68	0.39	0.06	0.37	4.00	False	False
’min cluster size’: 20, ’min samples’: 35	0.41	0.41	0.37	0.13	nan	2.00	False	False
’min cluster size’: 30, ’min samples’: 5	0.62	0.63	0.47	0.12	0.42	3.00	True	True
’min cluster size’: 30, ’min samples’: 10	0.68	0.69	0.44	0.08	0.29	4.00	False	False
’min cluster size’: 30, ’min samples’: 15	0.69	0.69	0.44	0.12	0.39	3.00	True	True
’min cluster size’: 30, ’min samples’: 20	0.69	0.69	0.42	0.08	0.37	4.00	False	False
’min cluster size’: 30, ’min samples’: 25	0.70	0.70	0.40	0.07	0.51	4.00	False	False
’min cluster size’: 30, ’min samples’: 30	0.68	0.68	0.39	0.06	0.37	4.00	False	False
’min cluster size’: 30, ’min samples’: 35	0.41	0.41	0.37	0.13	nan	2.00	False	False
’min cluster size’: 40, ’min samples’: 5	0.62	0.63	0.47	0.12	0.42	3.00	False	False
’min cluster size’: 40, ’min samples’: 10	0.68	0.69	0.44	0.08	0.29	4.00	False	False
’min cluster size’: 40, ’min samples’: 15	0.69	0.69	0.44	0.12	0.39	3.00	False	False
’min cluster size’: 40, ’min samples’: 20	0.74	0.75	0.48	0.13	0.38	3.00	False	False
’min cluster size’: 40, ’min samples’: 25	0.76	0.76	0.49	0.14	0.39	3.00	True	True
’min cluster size’: 40, ’min samples’: 30	0.74	0.74	0.47	0.13	0.37	3.00	False	False
’min cluster size’: 40, ’min samples’: 35	-0.00	0.00	0.00	0.00	nan	1.00	False	False

Table C.39: Wine dataset - HDBSCAN clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’min cluster size’: 10, ’min samples’: 5	0.27	0.29	-0.19	0.08	0.20	4.00	False	False
’min cluster size’: 10, ’min samples’: 10	0.34	0.35	-0.09	0.07	0.24	4.00	False	False
’min cluster size’: 10, ’min samples’: 15	0.56	0.57	0.27	0.08	0.24	3.00	False	False
’min cluster size’: 10, ’min samples’: 20	0.58	0.59	0.33	0.09	nan	2.00	True	False
’min cluster size’: 10, ’min samples’: 25	0.22	0.22	0.10	0.09	nan	2.00	False	False
’min cluster size’: 10, ’min samples’: 30	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’min cluster size’: 10, ’min samples’: 35	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’min cluster size’: 20, ’min samples’: 5	0.70	0.70	0.36	0.03	0.00	4.00	True	False
’min cluster size’: 20, ’min samples’: 10	0.65	0.66	0.27	0.06	0.37	4.00	True	True
’min cluster size’: 20, ’min samples’: 15	0.56	0.57	0.27	0.08	0.24	3.00	False	False
’min cluster size’: 20, ’min samples’: 20	0.58	0.59	0.33	0.09	nan	2.00	False	False
’min cluster size’: 20, ’min samples’: 25	0.22	0.22	0.10	0.09	nan	2.00	False	False
’min cluster size’: 20, ’min samples’: 30	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’min cluster size’: 20, ’min samples’: 35	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’min cluster size’: 30, ’min samples’: 5	0.70	0.70	0.36	0.03	0.00	4.00	False	False
’min cluster size’: 30, ’min samples’: 10	0.65	0.66	0.27	0.06	0.37	4.00	False	False
’min cluster size’: 30, ’min samples’: 15	0.56	0.57	0.27	0.08	0.24	3.00	False	False
’min cluster size’: 30, ’min samples’: 20	0.58	0.59	0.33	0.09	nan	2.00	False	False
’min cluster size’: 30, ’min samples’: 25	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’min cluster size’: 30, ’min samples’: 30	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’min cluster size’: 30, ’min samples’: 35	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’min cluster size’: 40, ’min samples’: 5	0.70	0.70	0.36	0.03	0.00	4.00	False	False
’min cluster size’: 40, ’min samples’: 10	0.71	0.71	0.29	0.08	0.35	3.00	True	True
’min cluster size’: 40, ’min samples’: 15	0.52	0.52	0.33	0.03	0.11	3.00	True	True
’min cluster size’: 40, ’min samples’: 20	0.58	0.59	0.33	0.09	nan	2.00	False	False
’min cluster size’: 40, ’min samples’: 25	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’min cluster size’: 40, ’min samples’: 30	-0.00	0.00	0.00	0.00	nan	1.00	False	False
’min cluster size’: 40, ’min samples’: 35	-0.00	0.00	0.00	0.00	nan	1.00	False	False

Table C.40: Wine dataset - HDBSCAN clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’min cluster size’: 5, ’min samples’: 5	0.33	0.36	-0.03	0.08	0.23	12.00	False	False
’min cluster size’: 5, ’min samples’: 10	0.35	0.36	-0.09	0.09	0.20	6.00	False	False
’min cluster size’: 5, ’min samples’: 15	0.61	0.62	0.29	0.09	0.41	5.00	False	False
’min cluster size’: 5, ’min samples’: 20	0.70	0.71	0.41	0.07	0.52	4.00	True	True
’min cluster size’: 5, ’min samples’: 25	0.70	0.70	0.41	0.09	0.35	4.00	False	False
’min cluster size’: 5, ’min samples’: 30	0.74	0.74	0.50	0.14	0.18	3.00	False	False
’min cluster size’: 5, ’min samples’: 35	0.56	0.56	0.42	0.13	nan	2.00	False	False
’min cluster size’: 10, ’min samples’: 5	0.31	0.33	-0.11	0.07	0.23	6.00	False	False
’min cluster size’: 10, ’min samples’: 10	0.33	0.35	-0.10	0.08	0.25	5.00	False	False
’min cluster size’: 10, ’min samples’: 15	0.63	0.64	0.34	0.08	0.49	4.00	True	True
’min cluster size’: 10, ’min samples’: 20	0.70	0.71	0.41	0.07	0.52	4.00	False	False
’min cluster size’: 10, ’min samples’: 25	0.70	0.70	0.41	0.09	0.35	4.00	False	False
’min cluster size’: 10, ’min samples’: 30	0.74	0.74	0.50	0.14	0.18	3.00	False	False
’min cluster size’: 10, ’min samples’: 35	0.56	0.56	0.42	0.13	nan	2.00	False	False
’min cluster size’: 15, ’min samples’: 5	0.55	0.55	0.13	0.11	0.29	5.00	False	False
’min cluster size’: 15, ’min samples’: 10	0.58	0.59	0.17	0.12	0.20	5.00	False	False
’min cluster size’: 15, ’min samples’: 15	0.63	0.64	0.34	0.08	0.49	4.00	False	False
’min cluster size’: 15, ’min samples’: 20	0.70	0.71	0.41	0.07	0.52	4.00	False	False
’min cluster size’: 15, ’min samples’: 25	0.70	0.70	0.41	0.09	0.35	4.00	False	False
’min cluster size’: 30, ’min samples’: 5	0.63	0.63	0.39	0.13	0.40	3.00	False	False
’min cluster size’: 30, ’min samples’: 10	0.66	0.66	0.39	0.14	0.38	3.00	False	False
’min cluster size’: 30, ’min samples’: 15	0.68	0.69	0.44	0.14	0.42	3.00	False	False
’min cluster size’: 30, ’min samples’: 20	0.77	0.77	0.51	0.15	0.47	3.00	True	True
’min cluster size’: 30, ’min samples’: 25	0.70	0.70	0.41	0.09	0.35	4.00	False	False
’min cluster size’: 30, ’min samples’: 30	0.74	0.74	0.50	0.14	0.18	3.00	False	False
’min cluster size’: 30, ’min samples’: 35	0.56	0.56	0.42	0.13	nan	2.00	False	False

Table C.41: Wine dataset - HDBSCAN clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’min cluster size’: 10, ’min samples’: 5	0.42	0.44	0.11	0.09	0.31	10.00	False	False
’min cluster size’: 10, ’min samples’: 10	0.44	0.45	0.19	0.06	0.24	8.00	False	False
’min cluster size’: 10, ’min samples’: 15	0.37	0.38	0.08	0.10	0.36	4.00	False	False
’min cluster size’: 10, ’min samples’: 20	0.58	0.59	0.31	0.07	0.45	4.00	False	False
’min cluster size’: 10, ’min samples’: 25	0.82	0.82	0.62	0.19	-0.13	3.00	True	True
’min cluster size’: 10, ’min samples’: 30	0.79	0.79	0.52	0.09	0.30	4.00	False	False
’min cluster size’: 10, ’min samples’: 35	0.73	0.73	0.50	0.09	0.30	4.00	False	False
’min cluster size’: 20, ’min samples’: 5	0.34	0.35	-0.00	0.05	0.33	4.00	False	False
’min cluster size’: 20, ’min samples’: 10	0.38	0.39	0.07	0.05	0.36	4.00	False	False
’min cluster size’: 20, ’min samples’: 15	0.55	0.56	0.31	0.08	0.46	4.00	False	False
’min cluster size’: 20, ’min samples’: 20	0.58	0.59	0.31	0.07	0.45	4.00	False	False
’min cluster size’: 20, ’min samples’: 25	0.82	0.82	0.62	0.19	-0.13	3.00	False	False
’min cluster size’: 20, ’min samples’: 30	0.79	0.79	0.52	0.09	0.30	4.00	False	False
’min cluster size’: 20, ’min samples’: 35	0.73	0.73	0.50	0.09	0.30	4.00	False	False
’min cluster size’: 30, ’min samples’: 5	0.46	0.47	0.25	0.06	0.44	4.00	False	False
’min cluster size’: 30, ’min samples’: 10	0.58	0.59	0.35	0.07	0.42	4.00	False	False
’min cluster size’: 30, ’min samples’: 15	0.70	0.71	0.53	0.12	0.50	4.00	True	True
’min cluster size’: 30, ’min samples’: 20	0.58	0.59	0.31	0.07	0.45	4.00	False	False
’min cluster size’: 30, ’min samples’: 25	0.82	0.82	0.62	0.19	-0.13	3.00	False	False
’min cluster size’: 30, ’min samples’: 30	0.79	0.79	0.52	0.09	0.30	4.00	False	False
’min cluster size’: 30, ’min samples’: 35	0.73	0.73	0.50	0.09	0.30	4.00	False	False
’min cluster size’: 40, ’min samples’: 5	0.64	0.65	0.49	0.10	0.42	4.00	False	False
’min cluster size’: 40, ’min samples’: 10	0.78	0.79	0.45	0.19	0.49	4.00	True	True
’min cluster size’: 40, ’min samples’: 15	0.79	0.80	0.48	0.18	0.52	4.00	True	True
’min cluster size’: 40, ’min samples’: 20	0.81	0.81	0.44	0.18	-0.06	4.00	False	False
’min cluster size’: 40, ’min samples’: 25	0.82	0.82	0.62	0.19	-0.13	3.00	False	False
’min cluster size’: 40, ’min samples’: 30	0.79	0.79	0.52	0.09	0.30	4.00	False	False
’min cluster size’: 40, ’min samples’: 35	0.73	0.73	0.50	0.09	0.30	4.00	False	False

Table C.42: Wine dataset - HDBSCAN clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’min cluster size’: 10, ’min samples’: 5	0.45	0.46	-0.02	0.04	0.33	5.00	False	False
’min cluster size’: 10, ’min samples’: 10	0.30	0.31	-0.24	0.02	0.25	4.00	False	False
’min cluster size’: 10, ’min samples’: 15	0.74	0.74	0.52	0.13	0.05	4.00	True	True
’min cluster size’: 10, ’min samples’: 20	0.80	0.80	0.57	0.01	0.11	4.00	False	False
’min cluster size’: 10, ’min samples’: 25	0.80	0.80	0.58	0.05	0.06	4.00	True	True
’min cluster size’: 10, ’min samples’: 30	0.77	0.78	0.50	0.08	-0.03	4.00	False	False
’min cluster size’: 10, ’min samples’: 35	0.77	0.77	0.57	0.02	0.13	4.00	False	False
’min cluster size’: 20, ’min samples’: 5	0.74	0.74	0.51	0.07	0.40	4.00	True	True
’min cluster size’: 20, ’min samples’: 10	0.75	0.76	0.47	0.05	0.48	4.00	True	True
’min cluster size’: 20, ’min samples’: 15	0.74	0.74	0.52	0.13	0.05	4.00	False	False
’min cluster size’: 20, ’min samples’: 20	0.80	0.80	0.57	0.01	0.11	4.00	False	False
’min cluster size’: 20, ’min samples’: 25	0.80	0.80	0.58	0.05	0.06	4.00	False	False
’min cluster size’: 20, ’min samples’: 30	0.77	0.78	0.50	0.08	-0.03	4.00	False	False
’min cluster size’: 20, ’min samples’: 35	0.77	0.77	0.57	0.02	0.13	4.00	False	False
’min cluster size’: 30, ’min samples’: 5	0.74	0.74	0.51	0.07	0.40	4.00	False	False
’min cluster size’: 30, ’min samples’: 10	0.81	0.81	0.57	0.06	-0.08	4.00	False	False
’min cluster size’: 30, ’min samples’: 15	0.81	0.81	0.58	0.06	0.04	4.00	True	True
’min cluster size’: 30, ’min samples’: 20	0.80	0.80	0.57	0.01	0.11	4.00	False	False
’min cluster size’: 30, ’min samples’: 25	0.80	0.80	0.58	0.05	0.06	4.00	False	False
’min cluster size’: 30, ’min samples’: 30	0.77	0.78	0.50	0.08	-0.03	4.00	False	False
’min cluster size’: 30, ’min samples’: 35	0.77	0.77	0.57	0.02	0.13	4.00	False	False
’min cluster size’: 40, ’min samples’: 5	0.79	0.79	0.58	0.04	0.42	4.00	True	True
’min cluster size’: 40, ’min samples’: 10	0.81	0.81	0.57	0.06	-0.08	4.00	False	False
’min cluster size’: 40, ’min samples’: 15	0.81	0.81	0.58	0.06	0.04	4.00	False	False
’min cluster size’: 40, ’min samples’: 20	0.80	0.80	0.57	0.01	0.11	4.00	False	False
’min cluster size’: 40, ’min samples’: 25	0.80	0.80	0.58	0.05	0.06	4.00	False	False
’min cluster size’: 40, ’min samples’: 30	0.77	0.78	0.50	0.08	-0.03	4.00	False	False
’min cluster size’: 40, ’min samples’: 35	0.77	0.77	0.57	0.02	0.13	4.00	False	False

Table C.43: Synthetic dataset - Hierarchical clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’linkage’: ’ward’, ’n clusters’: 2	0.59	0.59	0.59	0.15	0.06	2.00	True	True
’linkage’: ’ward’, ’n clusters’: 3	0.73	0.73	0.57	0.20	-0.09	3.00	False	False
’linkage’: ’ward’, ’n clusters’: 4	0.87	0.87	0.52	0.15	-0.07	4.00	False	False
’linkage’: ’ward’, ’n clusters’: 5	0.95	0.96	0.62	0.19	-0.04	5.00	True	True
’linkage’: ’ward’, ’n clusters’: 6	0.92	0.92	0.54	0.13	-0.10	6.00	False	False
’linkage’: ’ward’, ’n clusters’: 7	0.89	0.89	0.49	0.11	-0.21	7.00	False	False
’linkage’: ’ward’, ’n clusters’: 8	0.86	0.86	0.44	0.10	-0.40	8.00	False	False
’linkage’: ’ward’, ’n clusters’: 9	0.83	0.83	0.35	0.09	-0.59	9.00	False	False
’linkage’: ’ward’, ’n clusters’: 10	0.81	0.81	0.36	0.09	-0.58	10.00	False	False
’linkage’: ’single’, ’n clusters’: 2	-0.00	0.00	0.17	0.16	0.00	2.00	False	False
’linkage’: ’single’, ’n clusters’: 3	-0.00	0.00	-0.06	0.17	0.00	3.00	False	False
’linkage’: ’single’, ’n clusters’: 4	0.59	0.59	0.31	0.18	0.00	4.00	False	False
’linkage’: ’single’, ’n clusters’: 5	0.59	0.59	0.27	0.18	0.00	5.00	False	False
’linkage’: ’single’, ’n clusters’: 6	0.59	0.59	0.07	0.16	0.00	6.00	False	False
’linkage’: ’single’, ’n clusters’: 7	0.59	0.59	0.04	0.16	0.00	7.00	False	False
’linkage’: ’single’, ’n clusters’: 8	0.59	0.59	-0.06	0.15	0.00	8.00	False	False
’linkage’: ’single’, ’n clusters’: 9	0.58	0.59	-0.12	0.12	0.00	9.00	False	False
’linkage’: ’single’, ’n clusters’: 10	0.58	0.59	-0.12	0.13	0.00	10.00	False	False
’linkage’: ’complete’, ’n clusters’: 2	0.59	0.59	0.59	0.15	0.06	2.00	False	False
’linkage’: ’complete’, ’n clusters’: 3	0.73	0.73	0.57	0.20	-0.26	3.00	False	False
’linkage’: ’complete’, ’n clusters’: 4	0.80	0.80	0.45	0.14	-0.60	4.00	False	False
’linkage’: ’complete’, ’n clusters’: 5	0.89	0.89	0.57	0.18	-0.57	5.00	False	False
’linkage’: ’complete’, ’n clusters’: 6	0.86	0.86	0.50	0.13	-0.63	6.00	False	False
’linkage’: ’complete’, ’n clusters’: 7	0.83	0.84	0.44	0.11	-0.62	7.00	False	False
’linkage’: ’complete’, ’n clusters’: 8	0.81	0.81	0.36	0.10	-0.66	8.00	False	False
’linkage’: ’complete’, ’n clusters’: 9	0.79	0.79	0.36	0.10	-0.65	9.00	False	False
’linkage’: ’complete’, ’n clusters’: 10	0.76	0.76	0.29	0.09	-0.72	10.00	False	False
’linkage’: ’average’, ’n clusters’: 2	0.59	0.59	0.59	0.15	0.06	2.00	False	False
’linkage’: ’average’, ’n clusters’: 3	0.72	0.73	0.57	0.20	-0.09	3.00	False	False
’linkage’: ’average’, ’n clusters’: 4	0.87	0.87	0.52	0.15	-0.07	4.00	False	False
’linkage’: ’average’, ’n clusters’: 5	0.95	0.95	0.61	0.19	-0.04	5.00	False	False
’linkage’: ’average’, ’n clusters’: 6	0.95	0.95	0.60	0.21	0.00	6.00	True	False
’linkage’: ’average’, ’n clusters’: 7	0.95	0.95	0.57	0.22	0.00	7.00	True	False
’linkage’: ’average’, ’n clusters’: 8	0.95	0.95	0.55	0.21	0.00	8.00	False	False
’linkage’: ’average’, ’n clusters’: 9	0.94	0.94	0.50	0.19	0.00	9.00	False	False
’linkage’: ’average’, ’n clusters’: 10	0.94	0.94	0.45	0.19	0.00	10.00	False	False

Table C.44: Synthetic dataset - Hierarchical clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’linkage’: ’ward’, ’n clusters’: 2	0.59	0.59	0.59	0.15	0.06	2.00	True	True
’linkage’: ’ward’, ’n clusters’: 3	0.73	0.73	0.57	0.20	-0.09	3.00	False	False
’linkage’: ’ward’, ’n clusters’: 4	0.87	0.87	0.52	0.15	-0.07	4.00	False	False
’linkage’: ’ward’, ’n clusters’: 5	0.95	0.96	0.62	0.19	-0.04	5.00	True	True
’linkage’: ’ward’, ’n clusters’: 6	0.92	0.92	0.54	0.13	-0.10	6.00	False	False
’linkage’: ’ward’, ’n clusters’: 7	0.89	0.89	0.49	0.11	-0.21	7.00	False	False
’linkage’: ’ward’, ’n clusters’: 8	0.86	0.86	0.44	0.10	-0.40	8.00	False	False
’linkage’: ’ward’, ’n clusters’: 9	0.83	0.83	0.35	0.09	-0.59	9.00	False	False
’linkage’: ’ward’, ’n clusters’: 10	0.81	0.81	0.36	0.09	-0.58	10.00	False	False
’linkage’: ’single’, ’n clusters’: 2	-0.00	0.00	0.17	0.16	0.00	2.00	False	False
’linkage’: ’single’, ’n clusters’: 3	-0.00	0.00	-0.06	0.17	0.00	3.00	False	False
’linkage’: ’single’, ’n clusters’: 4	0.59	0.59	0.31	0.18	0.00	4.00	False	False
’linkage’: ’single’, ’n clusters’: 5	0.59	0.59	0.27	0.18	0.00	5.00	False	False
’linkage’: ’single’, ’n clusters’: 6	0.59	0.59	0.07	0.16	0.00	6.00	False	False
’linkage’: ’single’, ’n clusters’: 7	0.59	0.59	0.04	0.16	0.00	7.00	False	False
’linkage’: ’single’, ’n clusters’: 8	0.59	0.59	-0.06	0.15	0.00	8.00	False	False
’linkage’: ’single’, ’n clusters’: 9	0.58	0.59	-0.12	0.12	0.00	9.00	False	False
’linkage’: ’single’, ’n clusters’: 10	0.58	0.59	-0.12	0.13	0.00	10.00	False	False
’linkage’: ’complete’, ’n clusters’: 2	0.59	0.59	0.59	0.15	0.06	2.00	False	False
’linkage’: ’complete’, ’n clusters’: 3	0.73	0.73	0.57	0.20	-0.26	3.00	False	False
’linkage’: ’complete’, ’n clusters’: 4	0.80	0.80	0.45	0.14	-0.60	4.00	False	False
’linkage’: ’complete’, ’n clusters’: 5	0.89	0.89	0.57	0.18	-0.57	5.00	False	False
’linkage’: ’complete’, ’n clusters’: 6	0.86	0.86	0.50	0.13	-0.63	6.00	False	False
’linkage’: ’complete’, ’n clusters’: 7	0.83	0.84	0.44	0.11	-0.62	7.00	False	False
’linkage’: ’complete’, ’n clusters’: 8	0.81	0.81	0.36	0.10	-0.66	8.00	False	False
’linkage’: ’complete’, ’n clusters’: 9	0.79	0.79	0.36	0.10	-0.65	9.00	False	False
’linkage’: ’complete’, ’n clusters’: 10	0.76	0.76	0.29	0.09	-0.72	10.00	False	False
’linkage’: ’average’, ’n clusters’: 2	0.59	0.59	0.59	0.15	0.06	2.00	False	False
’linkage’: ’average’, ’n clusters’: 3	0.72	0.73	0.57	0.20	-0.09	3.00	False	False
’linkage’: ’average’, ’n clusters’: 4	0.87	0.87	0.52	0.15	-0.07	4.00	False	False
’linkage’: ’average’, ’n clusters’: 5	0.95	0.95	0.61	0.19	-0.04	5.00	False	False
’linkage’: ’average’, ’n clusters’: 6	0.95	0.95	0.60	0.21	0.00	6.00	True	False

Table C.45: Synthetic dataset - Hierarchical clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’linkage’: ’ward’, ’n clusters’: 2	0.59	0.59	0.59	0.15	0.06	2.00	True	True
’linkage’: ’ward’, ’n clusters’: 3	0.73	0.73	0.57	0.20	-0.09	3.00	False	False
’linkage’: ’ward’, ’n clusters’: 4	0.87	0.87	0.52	0.15	-0.07	4.00	False	False
’linkage’: ’ward’, ’n clusters’: 5	0.95	0.96	0.62	0.19	-0.04	5.00	True	True
’linkage’: ’ward’, ’n clusters’: 6	0.92	0.92	0.54	0.13	-0.10	6.00	False	False
’linkage’: ’ward’, ’n clusters’: 7	0.89	0.89	0.49	0.11	-0.21	7.00	False	False
’linkage’: ’ward’, ’n clusters’: 8	0.85	0.85	0.44	0.10	-0.42	8.00	False	False
’linkage’: ’ward’, ’n clusters’: 9	0.83	0.83	0.35	0.09	-0.62	9.00	False	False
’linkage’: ’ward’, ’n clusters’: 10	0.81	0.81	0.36	0.09	-0.61	10.00	False	False
’linkage’: ’single’, ’n clusters’: 2	-0.00	0.00	0.17	0.16	0.00	2.00	False	False
’linkage’: ’single’, ’n clusters’: 3	-0.00	0.00	-0.06	0.17	0.00	3.00	False	False
’linkage’: ’single’, ’n clusters’: 4	0.59	0.59	0.31	0.18	0.00	4.00	False	False
’linkage’: ’single’, ’n clusters’: 5	0.59	0.59	0.27	0.18	0.00	5.00	False	False
’linkage’: ’single’, ’n clusters’: 6	0.59	0.59	0.07	0.16	0.00	6.00	False	False
’linkage’: ’single’, ’n clusters’: 7	0.59	0.59	0.04	0.16	0.00	7.00	False	False
’linkage’: ’single’, ’n clusters’: 8	0.59	0.59	-0.06	0.15	0.00	8.00	False	False
’linkage’: ’single’, ’n clusters’: 9	0.58	0.59	-0.07	0.15	0.00	9.00	False	False
’linkage’: ’single’, ’n clusters’: 10	0.58	0.59	-0.13	0.15	0.00	10.00	False	False
’linkage’: ’complete’, ’n clusters’: 2	0.59	0.59	0.59	0.15	0.06	2.00	False	False
’linkage’: ’complete’, ’n clusters’: 3	0.73	0.73	0.57	0.20	-0.26	3.00	False	False
’linkage’: ’complete’, ’n clusters’: 4	0.80	0.80	0.45	0.14	-0.60	4.00	False	False
’linkage’: ’complete’, ’n clusters’: 5	0.89	0.89	0.57	0.18	-0.57	5.00	False	False
’linkage’: ’complete’, ’n clusters’: 6	0.86	0.86	0.50	0.13	-0.63	6.00	False	False
’linkage’: ’complete’, ’n clusters’: 7	0.83	0.84	0.44	0.11	-0.62	7.00	False	False
’linkage’: ’complete’, ’n clusters’: 8	0.81	0.81	0.36	0.10	-0.66	8.00	False	False
’linkage’: ’complete’, ’n clusters’: 9	0.79	0.79	0.36	0.10	-0.64	9.00	False	False
’linkage’: ’complete’, ’n clusters’: 10	0.76	0.76	0.29	0.09	-0.72	10.00	False	False
’linkage’: ’average’, ’n clusters’: 2	0.59	0.59	0.59	0.15	0.06	2.00	False	False
’linkage’: ’average’, ’n clusters’: 3	0.72	0.73	0.57	0.20	-0.09	3.00	False	False
’linkage’: ’average’, ’n clusters’: 4	0.87	0.87	0.52	0.15	-0.07	4.00	False	False
’linkage’: ’average’, ’n clusters’: 5	0.95	0.95	0.61	0.19	-0.04	5.00	False	False
’linkage’: ’average’, ’n clusters’: 6	0.95	0.95	0.60	0.21	0.00	6.00	True	False
’linkage’: ’average’, ’n clusters’: 7	0.95	0.95	0.57	0.22	0.00	7.00	True	False

Table C.46: Synthetic dataset - Hierarchical clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'linkage': 'ward', 'n clusters': 2	0.59	0.59	0.59	0.15	0.10	2.00	True	True
'linkage': 'ward', 'n clusters': 3	0.74	0.74	0.59	0.22	0.05	3.00	True	True
'linkage': 'ward', 'n clusters': 4	0.88	0.88	0.54	0.15	-0.12	4.00	False	False
'linkage': 'ward', 'n clusters': 5	0.96	0.96	0.64	0.21	0.07	5.00	True	True
'linkage': 'ward', 'n clusters': 6	0.93	0.93	0.57	0.14	-0.15	6.00	False	False
'linkage': 'ward', 'n clusters': 7	0.89	0.89	0.50	0.11	-0.25	7.00	False	False
'linkage': 'ward', 'n clusters': 8	0.86	0.86	0.42	0.10	-0.41	8.00	False	False
'linkage': 'ward', 'n clusters': 9	0.84	0.84	0.36	0.09	-0.56	9.00	False	False
'linkage': 'ward', 'n clusters': 10	0.81	0.81	0.30	0.08	-0.72	10.00	False	False
'linkage': 'single', 'n clusters': 2	-0.00	0.00	0.15	0.16	0.00	2.00	False	False
'linkage': 'single', 'n clusters': 3	-0.00	0.00	-0.11	0.16	0.00	3.00	False	False
'linkage': 'single', 'n clusters': 4	0.59	0.59	0.28	0.17	0.00	4.00	False	False
'linkage': 'single', 'n clusters': 5	0.59	0.59	0.09	0.15	0.00	5.00	False	False
'linkage': 'single', 'n clusters': 6	0.59	0.59	0.06	0.15	0.00	6.00	False	False
'linkage': 'single', 'n clusters': 7	0.59	0.59	0.00	0.12	0.00	7.00	False	False
'linkage': 'single', 'n clusters': 8	0.59	0.59	-0.12	0.11	0.00	8.00	False	False
'linkage': 'single', 'n clusters': 9	0.58	0.59	-0.12	0.11	0.00	9.00	False	False
'linkage': 'single', 'n clusters': 10	0.58	0.59	-0.14	0.12	0.00	10.00	False	False
'linkage': 'complete', 'n clusters': 2	0.59	0.59	0.59	0.15	0.10	2.00	False	False
'linkage': 'complete', 'n clusters': 3	0.73	0.73	0.59	0.22	0.01	3.00	False	False
'linkage': 'complete', 'n clusters': 4	0.87	0.87	0.51	0.14	-0.12	4.00	False	False
'linkage': 'complete', 'n clusters': 5	0.92	0.92	0.62	0.20	-0.21	5.00	False	False
'linkage': 'complete', 'n clusters': 6	0.89	0.89	0.55	0.14	-0.23	6.00	False	False
'linkage': 'complete', 'n clusters': 7	0.86	0.86	0.46	0.11	-0.47	7.00	False	False
'linkage': 'complete', 'n clusters': 8	0.86	0.86	0.45	0.11	-0.47	8.00	False	False
'linkage': 'complete', 'n clusters': 9	0.83	0.83	0.38	0.10	-0.54	9.00	False	False
'linkage': 'complete', 'n clusters': 10	0.81	0.81	0.38	0.10	-0.51	10.00	False	False
'linkage': 'average', 'n clusters': 2	0.59	0.59	0.59	0.15	0.10	2.00	True	True
'linkage': 'average', 'n clusters': 3	0.72	0.72	0.59	0.21	0.01	3.00	False	False
'linkage': 'average', 'n clusters': 4	0.86	0.86	0.54	0.15	-0.16	4.00	False	False
'linkage': 'average', 'n clusters': 5	0.94	0.94	0.63	0.20	-0.23	5.00	False	False
'linkage': 'average', 'n clusters': 6	0.94	0.94	0.62	0.22	0.00	6.00	True	False
'linkage': 'average', 'n clusters': 7	0.94	0.94	0.58	0.23	0.00	7.00	True	False

Table C.47: Synthetic dataset - Hierarchical clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’linkage’: ’ward’, ’n clusters’: 2	0.59	0.59	0.52	0.12	0.16	2.00	False	False
’linkage’: ’ward’, ’n clusters’: 3	0.74	0.74	0.58	0.19	0.47	3.00	False	False
’linkage’: ’ward’, ’n clusters’: 4	0.88	0.88	0.61	0.17	0.59	4.00	False	False
’linkage’: ’ward’, ’n clusters’: 5	0.97	0.97	0.75	0.28	0.78	5.00	True	True
’linkage’: ’ward’, ’n clusters’: 6	0.93	0.93	0.67	0.17	0.55	6.00	False	False
’linkage’: ’ward’, ’n clusters’: 7	0.89	0.89	0.59	0.13	0.36	7.00	False	False
’linkage’: ’ward’, ’n clusters’: 8	0.86	0.86	0.52	0.11	0.13	8.00	False	False
’linkage’: ’ward’, ’n clusters’: 9	0.83	0.83	0.44	0.10	-0.08	9.00	False	False
’linkage’: ’ward’, ’n clusters’: 10	0.80	0.80	0.37	0.09	-0.28	10.00	False	False
’linkage’: ’single’, ’n clusters’: 2	0.47	0.47	0.33	0.12	0.28	2.00	False	False
’linkage’: ’single’, ’n clusters’: 3	0.74	0.74	0.58	0.19	0.47	3.00	False	False
’linkage’: ’single’, ’n clusters’: 4	0.88	0.88	0.61	0.17	0.59	4.00	False	False
’linkage’: ’single’, ’n clusters’: 5	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’linkage’: ’single’, ’n clusters’: 6	0.96	0.96	0.57	0.16	0.00	6.00	False	False
’linkage’: ’single’, ’n clusters’: 7	0.96	0.96	0.44	0.15	0.00	7.00	False	False
’linkage’: ’single’, ’n clusters’: 8	0.96	0.96	0.42	0.15	0.00	8.00	False	False
’linkage’: ’single’, ’n clusters’: 9	0.96	0.96	0.29	0.14	0.00	9.00	False	False
’linkage’: ’single’, ’n clusters’: 10	0.96	0.96	0.23	0.12	0.00	10.00	False	False
’linkage’: ’complete’, ’n clusters’: 2	0.59	0.59	0.52	0.12	0.16	2.00	False	False
’linkage’: ’complete’, ’n clusters’: 3	0.74	0.74	0.58	0.19	0.47	3.00	False	False
’linkage’: ’complete’, ’n clusters’: 4	0.88	0.88	0.61	0.17	0.59	4.00	False	False
’linkage’: ’complete’, ’n clusters’: 5	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’linkage’: ’complete’, ’n clusters’: 6	0.93	0.93	0.66	0.16	0.58	6.00	False	False
’linkage’: ’complete’, ’n clusters’: 7	0.89	0.89	0.59	0.13	0.35	7.00	False	False
’linkage’: ’complete’, ’n clusters’: 8	0.86	0.86	0.51	0.11	0.12	8.00	False	False
’linkage’: ’complete’, ’n clusters’: 9	0.83	0.83	0.44	0.10	-0.10	9.00	False	False
’linkage’: ’complete’, ’n clusters’: 10	0.80	0.80	0.36	0.09	-0.35	10.00	False	False
’linkage’: ’average’, ’n clusters’: 2	0.59	0.59	0.52	0.12	0.16	2.00	False	False
’linkage’: ’average’, ’n clusters’: 3	0.74	0.74	0.58	0.19	0.47	3.00	False	False
’linkage’: ’average’, ’n clusters’: 4	0.88	0.88	0.61	0.17	0.59	4.00	False	False
’linkage’: ’average’, ’n clusters’: 5	0.97	0.97	0.75	0.28	0.78	5.00	False	False

Table C.48: Synthetic dataset - Hierarchical clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’linkage’: ’ward’, ’n clusters’: 2	0.59	0.59	0.73	0.21	0.55	2.00	False	False
’linkage’: ’ward’, ’n clusters’: 3	0.77	0.77	0.70	0.22	0.57	3.00	False	False
’linkage’: ’ward’, ’n clusters’: 4	0.88	0.88	0.75	0.29	0.72	4.00	False	False
’linkage’: ’ward’, ’n clusters’: 5	0.97	0.97	0.85	0.43	0.91	5.00	True	True
’linkage’: ’ward’, ’n clusters’: 6	0.93	0.93	0.76	0.21	0.75	6.00	False	False
’linkage’: ’ward’, ’n clusters’: 7	0.89	0.89	0.68	0.16	0.49	7.00	False	False
’linkage’: ’ward’, ’n clusters’: 8	0.86	0.86	0.58	0.12	0.26	8.00	False	False
’linkage’: ’ward’, ’n clusters’: 9	0.83	0.83	0.50	0.11	0.02	9.00	False	False
’linkage’: ’ward’, ’n clusters’: 10	0.80	0.80	0.40	0.09	-0.21	10.00	False	False
’linkage’: ’single’, ’n clusters’: 2	0.59	0.59	0.73	0.21	0.55	2.00	False	False
’linkage’: ’single’, ’n clusters’: 3	0.77	0.77	0.70	0.22	0.57	3.00	False	False
’linkage’: ’single’, ’n clusters’: 4	0.88	0.88	0.75	0.29	0.72	4.00	False	False
’linkage’: ’single’, ’n clusters’: 5	0.97	0.97	0.85	0.43	0.91	5.00	False	False
’linkage’: ’single’, ’n clusters’: 6	0.97	0.97	0.68	0.23	0.00	6.00	False	False
’linkage’: ’single’, ’n clusters’: 7	0.96	0.96	0.61	0.17	0.00	7.00	False	False
’linkage’: ’single’, ’n clusters’: 8	0.96	0.96	0.60	0.15	0.00	8.00	False	False
’linkage’: ’single’, ’n clusters’: 9	0.96	0.96	0.42	0.11	0.00	9.00	False	False
’linkage’: ’single’, ’n clusters’: 10	0.94	0.94	0.29	0.11	0.00	10.00	False	False
’linkage’: ’complete’, ’n clusters’: 2	0.59	0.59	0.73	0.21	0.55	2.00	False	False
’linkage’: ’complete’, ’n clusters’: 3	0.77	0.77	0.70	0.22	0.57	3.00	False	False
’linkage’: ’complete’, ’n clusters’: 4	0.88	0.88	0.75	0.29	0.72	4.00	False	False
’linkage’: ’complete’, ’n clusters’: 5	0.97	0.97	0.85	0.43	0.91	5.00	False	False

Table C.49: Synthetic dataset - *k*-means clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.59	0.59	0.59	0.15	0.17	2.00	True	True
’n clusters’: 3	0.72	0.72	0.57	0.20	-0.47	3.00	True	True
’n clusters’: 4	0.87	0.87	0.53	0.15	-0.38	4.00	False	False
’n clusters’: 5	0.97	0.97	0.62	0.19	-0.03	5.00	True	True
’n clusters’: 6	0.92	0.92	0.55	0.13	-0.09	6.00	False	False
’n clusters’: 7	0.88	0.89	0.47	0.11	-0.25	7.00	False	False
’n clusters’: 8	0.85	0.85	0.42	0.10	-0.72	8.00	False	False
’n clusters’: 9	0.82	0.82	0.37	0.09	-0.48	9.00	False	False
’n clusters’: 10	0.80	0.79	0.32	0.08	-0.71	10.00	False	False
’n clusters’: 11	0.77	0.78	0.32	0.09	-0.71	11.00	False	False

Table C.50: Synthetic dataset - *k*-means clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.59	0.59	0.59	0.15	0.17	2.00	True	True
’n clusters’: 3	0.72	0.72	0.57	0.20	-0.47	3.00	True	True
’n clusters’: 4	0.87	0.87	0.53	0.15	-0.38	4.00	False	False
’n clusters’: 5	0.97	0.97	0.62	0.19	-0.03	5.00	True	True
’n clusters’: 6	0.92	0.92	0.54	0.13	-0.09	6.00	False	False
’n clusters’: 7	0.88	0.89	0.46	0.11	-0.47	7.00	False	False
’n clusters’: 8	0.85	0.85	0.41	0.10	-0.38	8.00	False	False
’n clusters’: 9	0.81	0.82	0.36	0.09	-0.51	9.00	False	False
’n clusters’: 10	0.79	0.79	0.32	0.08	-0.72	10.00	False	False
’n clusters’: 11	0.77	0.78	0.32	0.09	-0.74	11.00	False	False

Table C.51: Synthetic dataset - *k*-means clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.59	0.59	0.59	0.15	0.17	2.00	True	True
’n clusters’: 3	0.72	0.72	0.57	0.20	-0.47	3.00	True	True
’n clusters’: 4	0.87	0.87	0.53	0.15	-0.38	4.00	False	False
’n clusters’: 5	0.97	0.97	0.62	0.19	-0.03	5.00	True	True
’n clusters’: 6	0.92	0.92	0.55	0.13	-0.10	6.00	False	False
’n clusters’: 7	0.88	0.88	0.48	0.11	-0.33	7.00	False	False
’n clusters’: 8	0.84	0.84	0.42	0.10	-0.38	8.00	False	False
’n clusters’: 9	0.82	0.82	0.36	0.09	-0.60	9.00	False	False
’n clusters’: 10	0.79	0.80	0.32	0.08	-0.74	10.00	False	False
’n clusters’: 11	0.77	0.78	0.32	0.09	-0.72	11.00	False	False

Table C.52: Synthetic dataset - *k*-means clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.59	0.59	0.59	0.15	0.10	2.00	True	True
’n clusters’: 3	0.72	0.72	0.59	0.21	-0.32	3.00	True	True
’n clusters’: 4	0.86	0.86	0.54	0.15	-0.03	4.00	False	False
’n clusters’: 5	0.97	0.97	0.64	0.21	0.06	5.00	True	True
’n clusters’: 6	0.93	0.92	0.57	0.14	-0.04	6.00	False	False
’n clusters’: 7	0.89	0.89	0.50	0.11	-0.35	7.00	False	False
’n clusters’: 8	0.85	0.85	0.47	0.10	-0.45	8.00	False	False
’n clusters’: 9	0.83	0.82	0.37	0.09	-0.57	9.00	False	False
’n clusters’: 10	0.80	0.79	0.33	0.08	-0.73	10.00	False	False
’n clusters’: 11	0.78	0.78	0.34	0.09	-0.72	11.00	False	False

Table C.53: Synthetic dataset - *k*-means clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.52	0.52	0.51	0.12	-0.81	2.00	False	False
’n clusters’: 3	0.72	0.72	0.57	0.19	-0.52	3.00	False	False
’n clusters’: 4	0.88	0.88	0.61	0.17	0.59	4.00	False	False
’n clusters’: 5	0.97	0.97	0.75	0.28	0.78	5.00	True	True
’n clusters’: 6	0.93	0.93	0.67	0.17	0.58	6.00	False	False
’n clusters’: 7	0.89	0.89	0.60	0.13	0.36	7.00	False	False
’n clusters’: 8	0.86	0.86	0.52	0.11	0.05	8.00	False	False
’n clusters’: 9	0.83	0.83	0.45	0.10	-0.13	9.00	False	False
’n clusters’: 10	0.80	0.80	0.38	0.09	-0.38	10.00	False	False
’n clusters’: 11	0.78	0.78	0.39	0.10	-0.39	11.00	False	False

Table C.54: Synthetic dataset - *k*-means clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.59	0.59	0.73	0.21	0.55	2.00	False	False
’n clusters’: 3	0.77	0.77	0.70	0.22	0.57	3.00	False	False
’n clusters’: 4	0.88	0.88	0.75	0.29	0.72	4.00	False	False
’n clusters’: 5	0.97	0.97	0.85	0.43	0.91	5.00	True	True
’n clusters’: 6	0.93	0.93	0.76	0.21	0.66	6.00	False	False
’n clusters’: 7	0.89	0.89	0.68	0.16	0.35	7.00	False	False
’n clusters’: 8	0.86	0.86	0.57	0.13	0.15	8.00	False	False
’n clusters’: 9	0.83	0.83	0.48	0.11	-0.11	9.00	False	False
’n clusters’: 10	0.80	0.80	0.39	0.09	-0.36	10.00	False	False
’n clusters’: 11	0.78	0.78	0.42	0.10	-0.47	11.00	False	False

Table C.55: Synthetic dataset - *k*-medoids clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.59	0.59	0.59	0.15	0.06	2.00	True	True
’n clusters’: 3	0.73	0.73	0.48	0.12	-0.23	3.00	False	False
’n clusters’: 4	0.85	0.85	0.51	0.14	-0.27	4.00	False	False
’n clusters’: 5	0.97	0.97	0.62	0.19	-0.03	5.00	True	True
’n clusters’: 6	0.93	0.93	0.55	0.13	-0.14	6.00	False	False
’n clusters’: 7	0.90	0.90	0.54	0.14	-0.18	7.00	False	False
’n clusters’: 8	0.86	0.87	0.50	0.12	-0.21	8.00	False	False
’n clusters’: 9	0.85	0.85	0.49	0.13	-0.22	9.00	False	False
’n clusters’: 10	0.83	0.83	0.48	0.12	-0.35	10.00	False	False
’n clusters’: 11	0.82	0.82	0.48	0.12	-0.24	11.00	False	False

Table C.56: Synthetic dataset - *k*-medoids clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.59	0.59	0.59	0.15	0.06	2.00	True	True
’n clusters’: 3	0.73	0.73	0.48	0.12	-0.23	3.00	False	False
’n clusters’: 4	0.85	0.85	0.51	0.14	-0.27	4.00	False	False
’n clusters’: 5	0.97	0.97	0.62	0.19	-0.03	5.00	True	True
’n clusters’: 6	0.93	0.93	0.55	0.13	-0.14	6.00	False	False
’n clusters’: 7	0.90	0.90	0.54	0.14	-0.18	7.00	False	False
’n clusters’: 8	0.86	0.87	0.50	0.12	-0.21	8.00	False	False
’n clusters’: 9	0.85	0.85	0.49	0.13	-0.22	9.00	False	False
’n clusters’: 10	0.83	0.83	0.48	0.12	-0.35	10.00	False	False
’n clusters’: 11	0.82	0.82	0.48	0.12	-0.24	11.00	False	False

Table C.57: Synthetic dataset - *k*-medoids clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'n clusters': 2	0.59	0.59	0.59	0.15	0.06	2.00	True	True
'n clusters': 3	0.73	0.73	0.48	0.12	-0.23	3.00	False	False
'n clusters': 4	0.85	0.85	0.51	0.14	-0.27	4.00	False	False
'n clusters': 5	0.97	0.97	0.62	0.19	-0.03	5.00	True	True
'n clusters': 6	0.93	0.93	0.55	0.13	-0.14	6.00	False	False
'n clusters': 7	0.90	0.90	0.54	0.14	-0.18	7.00	False	False
'n clusters': 8	0.86	0.87	0.50	0.12	-0.21	8.00	False	False
'n clusters': 9	0.85	0.85	0.49	0.13	-0.22	9.00	False	False
'n clusters': 10	0.83	0.83	0.48	0.12	-0.35	10.00	False	False
'n clusters': 11	0.82	0.82	0.48	0.12	-0.24	11.00	False	False

Table C.58: Synthetic dataset - *k*-medoids clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
n clusters': 2	0.59	0.59	0.59	0.15	0.10	2.00	True	True
'n clusters': 3	0.75	0.75	0.50	0.13	-0.09	3.00	False	False
'n clusters': 4	0.86	0.86	0.54	0.15	-0.03	4.00	False	False
'n clusters': 5	0.97	0.97	0.64	0.21	0.06	5.00	True	True
'n clusters': 6	0.92	0.92	0.57	0.14	-0.12	6.00	False	False
'n clusters': 7	0.88	0.88	0.52	0.12	-0.22	7.00	False	False
'n clusters': 8	0.87	0.87	0.51	0.12	-0.19	8.00	False	False
'n clusters': 9	0.73	0.74	0.39	0.11	-0.46	9.00	False	False
'n clusters': 10	0.84	0.84	0.49	0.12	-0.18	10.00	False	False
'n clusters': 11	0.83	0.83	0.47	0.12	-0.36	11.00	False	False

Table C.59: Synthetic dataset - *k*-medoids clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'n clusters': 2	0.49	0.49	0.49	0.12	-0.83	2.00	False	False
'n clusters': 3	0.70	0.70	0.43	0.10	-0.50	3.00	False	False
'n clusters': 4	0.87	0.87	0.59	0.16	0.46	4.00	False	False
'n clusters': 5	0.97	0.97	0.75	0.28	0.78	5.00	True	True
'n clusters': 6	0.93	0.93	0.65	0.15	0.54	6.00	False	False
'n clusters': 7	0.74	0.74	0.35	0.10	-0.44	7.00	False	False
'n clusters': 8	0.87	0.87	0.60	0.15	0.31	8.00	False	False
'n clusters': 9	0.86	0.86	0.59	0.14	0.33	9.00	False	False
'n clusters': 10	0.85	0.85	0.59	0.13	0.33	10.00	False	False
'n clusters': 11	0.87	0.87	0.64	0.15	0.59	11.00	False	False

Table C.60: Synthetic dataset - *k*-medoids clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'n clusters': 2	0.59	0.59	0.73	0.21	0.55	2.00	False	False
'n clusters': 3	0.77	0.77	0.58	0.14	0.23	3.00	False	False
'n clusters': 4	0.87	0.87	0.78	0.34	0.76	4.00	True	True
'n clusters': 5	0.84	0.84	0.70	0.18	0.51	5.00	False	False
'n clusters': 6	0.81	0.81	0.70	0.20	0.48	6.00	False	False
'n clusters': 7	0.80	0.80	0.70	0.18	0.46	7.00	False	False
'n clusters': 8	0.79	0.80	0.68	0.15	0.46	8.00	False	False
'n clusters': 9	0.79	0.79	0.69	0.16	0.49	9.00	False	False
'n clusters': 10	0.78	0.78	0.69	0.16	0.51	10.00	False	False
'n clusters': 11	0.78	0.78	0.68	0.15	0.49	11.00	False	False

Table C.61: Synthetic dataset - GMM clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’covariance type’: ’full’, ’n ’: 11	0.80	0.81	0.32	0.11	-0.50	11.00	False	False
’covariance type’: ’tied’, ’n ’: 2	0.58	0.58	0.59	0.15	-0.11	2.00	False	False
’covariance type’: ’tied’, ’n ’: 3	0.75	0.73	0.57	0.20	-0.18	3.00	True	True
’covariance type’: ’tied’, ’n ’: 4	0.86	0.86	0.54	0.15	-0.37	4.00	False	False
’covariance type’: ’tied’, ’n ’: 5	0.97	0.97	0.62	0.19	-0.27	5.00	False	False
’covariance type’: ’tied’, ’n ’: 6	0.93	0.93	0.54	0.13	-0.13	6.00	False	False
’covariance type’: ’tied’, ’n ’: 7	0.90	0.91	0.45	0.11	-0.28	7.00	False	False
’covariance type’: ’tied’, ’n ’: 8	0.87	0.89	0.42	0.11	-0.38	8.00	False	False
’covariance type’: ’tied’, ’n ’: 9	0.84	0.83	0.36	0.10	-0.63	9.00	False	False
’covariance type’: ’tied’, ’n ’: 10	0.82	0.81	0.36	0.09	-0.42	10.00	False	False
’covariance type’: ’tied’, ’n ’: 11	0.79	0.79	0.33	0.10	-0.75	11.00	False	False
’covariance type’: ’diag’, ’n ’: 2	0.59	0.59	0.59	0.15	0.06	2.00	True	True
’covariance type’: ’diag’, ’n ’: 3	0.75	0.69	0.49	0.20	0.04	3.00	True	True
’covariance type’: ’diag’, ’n ’: 4	0.86	0.87	0.52	0.13	-0.08	4.00	False	False
’covariance type’: ’diag’, ’n ’: 5	0.97	0.97	0.62	0.19	-0.03	5.00	True	True
’covariance type’: ’diag’, ’n ’: 6	0.93	0.94	0.54	0.13	-0.14	6.00	False	False
’covariance type’: ’diag’, ’n ’: 7	0.89	0.89	0.46	0.11	-0.32	7.00	False	False
’covariance type’: ’diag’, ’n ’: 8	0.88	0.86	0.42	0.10	-0.45	8.00	False	False
’covariance type’: ’diag’, ’n ’: 9	0.84	0.83	0.38	0.09	-0.56	9.00	False	False
’covariance type’: ’diag’, ’n ’: 10	0.81	0.81	0.36	0.09	-0.65	10.00	False	False
’covariance type’: ’diag’, ’n ’: 11	0.80	0.78	0.32	0.09	-0.65	11.00	False	False
’covariance type’: ’spherical’, ’n ’: 2	0.59	0.59	0.59	0.15	0.06	2.00	False	False
’covariance type’: ’spherical’, ’n ’: 3	0.73	0.69	0.45	0.13	-0.26	3.00	False	False
’covariance type’: ’spherical’, ’n ’: 4	0.81	0.83	0.56	0.15	0.03	4.00	False	False
’covariance type’: ’spherical’, ’n ’: 5	0.97	0.97	0.62	0.19	-0.03	5.00	True	True
’covariance type’: ’spherical’, ’n ’: 6	0.93	0.93	0.54	0.14	-0.14	6.00	False	False
’covariance type’: ’spherical’, ’n ’: 7	0.91	0.90	0.46	0.14	-0.32	7.00	False	False
’covariance type’: ’spherical’, ’n ’: 8	0.86	0.86	0.48	0.10	-0.48	8.00	False	False
’covariance type’: ’spherical’, ’n ’: 9	0.83	0.84	0.43	0.09	-0.61	9.00	False	False
’covariance type’: ’spherical’, ’n ’: 10	0.81	0.81	0.38	0.11	-0.72	10.00	False	False
’covariance type’: ’spherical’, ’n ’: 11	0.81	0.82	0.31	0.09	-0.63	11.00	False	False

Table C.62: Synthetic dataset - GMM clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’covariance type’: ’full’, ’n ’: 10	0.82	0.81	0.35	0.08	-0.61	10.00	False	False
’covariance type’: ’full’, ’n ’: 11	0.78	0.78	0.32	0.09	-0.68	11.00	False	False
’covariance type’: ’tied’, ’n ’: 2	0.58	0.58	0.59	0.15	-0.11	2.00	False	False
’covariance type’: ’tied’, ’n ’: 3	0.73	0.75	0.57	0.20	-0.27	3.00	True	True
’covariance type’: ’tied’, ’n ’: 4	0.83	0.84	0.51	0.15	-0.17	4.00	False	False
’covariance type’: ’tied’, ’n ’: 5	0.97	0.97	0.62	0.19	-0.03	5.00	False	False
’covariance type’: ’tied’, ’n ’: 6	0.93	0.93	0.53	0.13	-0.20	6.00	False	False
’covariance type’: ’tied’, ’n ’: 7	0.90	0.90	0.50	0.11	-0.32	7.00	False	False
’covariance type’: ’tied’, ’n ’: 8	0.87	0.89	0.39	0.10	-0.30	8.00	False	False
’covariance type’: ’tied’, ’n ’: 9	0.85	0.85	0.36	0.09	-0.55	9.00	False	False
’covariance type’: ’tied’, ’n ’: 10	0.81	0.81	0.31	0.11	-0.54	10.00	False	False
’covariance type’: ’tied’, ’n ’: 11	0.81	0.80	0.37	0.09	-0.49	11.00	False	False
’covariance type’: ’diag’, ’n ’: 2	0.58	0.58	0.59	0.15	0.06	2.00	False	False
’covariance type’: ’diag’, ’n ’: 3	0.73	0.73	0.49	0.13	-0.08	3.00	False	False
’covariance type’: ’diag’, ’n ’: 4	0.88	0.88	0.45	0.15	0.09	4.00	True	True
’covariance type’: ’diag’, ’n ’: 5	0.97	0.97	0.62	0.19	-0.03	5.00	True	True
’covariance type’: ’diag’, ’n ’: 6	0.93	0.93	0.55	0.13	-0.15	6.00	False	False
’covariance type’: ’diag’, ’n ’: 7	0.90	0.89	0.49	0.11	-0.42	7.00	False	False
’covariance type’: ’diag’, ’n ’: 8	0.86	0.86	0.41	0.12	-0.38	8.00	False	False
’covariance type’: ’diag’, ’n ’: 9	0.83	0.83	0.42	0.08	-0.54	9.00	False	False
’covariance type’: ’diag’, ’n ’: 10	0.81	0.83	0.35	0.09	-0.54	10.00	False	False
’covariance type’: ’diag’, ’n ’: 11	0.80	0.79	0.36	0.08	-0.60	11.00	False	False
’covariance type’: ’spherical’, ’n ’: 2	0.59	0.59	0.59	0.15	0.06	2.00	True	True
’covariance type’: ’spherical’, ’n ’: 3	0.68	0.73	0.48	0.20	-0.18	3.00	True	True
’covariance type’: ’spherical’, ’n ’: 4	0.83	0.83	0.48	0.18	-0.35	4.00	False	False
’covariance type’: ’spherical’, ’n ’: 5	0.97	0.97	0.62	0.19	-0.03	5.00	True	True
’covariance type’: ’spherical’, ’n ’: 6	0.93	0.93	0.54	0.13	-0.20	6.00	False	False
’covariance type’: ’spherical’, ’n ’: 7	0.89	0.91	0.45	0.11	-0.29	7.00	False	False
’covariance type’: ’spherical’, ’n ’: 8	0.86	0.86	0.41	0.09	-0.46	8.00	False	False
’covariance type’: ’spherical’, ’n ’: 9	0.83	0.83	0.36	0.09	-0.39	9.00	False	False
’covariance type’: ’spherical’, ’n ’: 10	0.82	0.82	0.39	0.09	-0.50	10.00	False	False
’covariance type’: ’spherical’, ’n ’: 11	0.78	0.80	0.36	0.10	-0.73	11.00	False	False

Table C.63: Synthetic dataset - GMM clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’covariance type’: ’full’, ’n ’: 2	0.58	0.59	0.59	0.15	0.06	2.00	False	False
’covariance type’: ’full’, ’n ’: 3	0.73	0.76	0.49	0.20	-0.08	3.00	True	True
’covariance type’: ’full’, ’n ’: 4	0.88	0.88	0.53	0.14	-0.07	4.00	False	False
’covariance type’: ’full’, ’n ’: 5	0.97	0.97	0.62	0.19	-0.03	5.00	False	False
’covariance type’: ’full’, ’n ’: 6	0.93	0.93	0.54	0.13	-0.14	6.00	False	False
’covariance type’: ’full’, ’n ’: 7	0.90	0.89	0.49	0.11	-0.37	7.00	False	False
’covariance type’: ’full’, ’n ’: 8	0.87	0.86	0.40	0.10	-0.35	8.00	False	False
’covariance type’: ’full’, ’n ’: 9	0.84	0.84	0.36	0.09	-0.61	9.00	False	False
’covariance type’: ’full’, ’n ’: 10	0.83	0.80	0.36	0.08	-0.54	10.00	False	False
’covariance type’: ’full’, ’n ’: 11	0.78	0.80	0.31	0.09	-0.69	11.00	False	False
’covariance type’: ’tied’, ’n ’: 2	0.58	0.58	0.59	0.15	-0.11	2.00	False	False
’covariance type’: ’tied’, ’n ’: 3	0.75	0.75	0.49	0.12	-0.18	3.00	False	False
’covariance type’: ’tied’, ’n ’: 4	0.86	0.86	0.52	0.15	-0.17	4.00	False	False
’covariance type’: ’tied’, ’n ’: 5	0.97	0.79	0.44	0.19	-0.03	5.00	False	False
’covariance type’: ’tied’, ’n ’: 6	0.93	0.93	0.53	0.13	-0.24	6.00	False	False
’covariance type’: ’tied’, ’n ’: 7	0.90	0.90	0.48	0.11	-0.27	7.00	False	False
’covariance type’: ’tied’, ’n ’: 8	0.88	0.86	0.42	0.10	-0.46	8.00	False	False
’covariance type’: ’tied’, ’n ’: 9	0.83	0.83	0.36	0.10	-0.46	9.00	False	False
’covariance type’: ’tied’, ’n ’: 10	0.80	0.82	0.34	0.09	-0.60	10.00	False	False
’covariance type’: ’tied’, ’n ’: 11	0.81	0.80	0.38	0.09	-0.62	11.00	False	False
’covariance type’: ’diag’, ’n ’: 2	0.59	0.59	0.59	0.15	0.06	2.00	True	True
’covariance type’: ’diag’, ’n ’: 3	0.73	0.73	0.57	0.12	0.05	3.00	False	False
’covariance type’: ’diag’, ’n ’: 4	0.81	0.87	0.56	0.15	-0.08	4.00	False	False
’covariance type’: ’diag’, ’n ’: 5	0.97	0.97	0.62	0.19	-0.03	5.00	True	True
’covariance type’: ’diag’, ’n ’: 6	0.95	0.94	0.53	0.13	-0.20	6.00	False	False
’covariance type’: ’diag’, ’n ’: 7	0.90	0.89	0.47	0.11	-0.17	7.00	False	False
’covariance type’: ’diag’, ’n ’: 8	0.88	0.88	0.39	0.09	-0.40	8.00	False	False

Table C.64: Synthetic dataset - GMM clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’covariance type’: ’full’, ’n ’: 2	0.58	0.59	0.59	0.15	0.10	2.00	True	True
’covariance type’: ’full’, ’n ’: 3	0.76	0.77	0.49	0.22	-0.05	3.00	True	True
’covariance type’: ’full’, ’n ’: 4	0.88	0.88	0.53	0.15	-0.03	4.00	False	False
’covariance type’: ’full’, ’n ’: 5	0.97	0.97	0.64	0.21	0.06	5.00	True	True
’covariance type’: ’full’, ’n ’: 6	0.93	0.93	0.57	0.14	-0.06	6.00	False	False
’covariance type’: ’full’, ’n ’: 7	0.89	0.89	0.52	0.11	-0.25	7.00	False	False
’covariance type’: ’full’, ’n ’: 8	0.87	0.86	0.52	0.09	-0.42	8.00	False	False
’covariance type’: ’full’, ’n ’: 9	0.84	0.83	0.42	0.09	-0.57	9.00	False	False
’covariance type’: ’full’, ’n ’: 10	0.81	0.81	0.38	0.10	-0.71	10.00	False	False
’covariance type’: ’full’, ’n ’: 11	0.80	0.79	0.36	0.09	-0.67	11.00	False	False
’covariance type’: ’tied’, ’n ’: 2	0.58	0.58	0.59	0.15	0.19	2.00	True	True
’covariance type’: ’tied’, ’n ’: 3	0.73	0.73	0.48	0.21	-0.22	3.00	False	False
’covariance type’: ’tied’, ’n ’: 4	0.87	0.85	0.53	0.15	-0.21	4.00	False	False
’covariance type’: ’tied’, ’n ’: 5	0.97	0.97	0.64	0.21	0.06	5.00	True	True
’covariance type’: ’tied’, ’n ’: 6	0.93	0.93	0.55	0.14	-0.05	6.00	False	False
’covariance type’: ’tied’, ’n ’: 7	0.90	0.90	0.43	0.11	-0.26	7.00	False	False
’covariance type’: ’tied’, ’n ’: 8	0.88	0.88	0.42	0.12	-0.34	8.00	False	False
’covariance type’: ’tied’, ’n ’: 9	0.84	0.84	0.40	0.11	-0.58	9.00	False	False
’covariance type’: ’tied’, ’n ’: 10	0.83	0.82	0.37	0.08	-0.67	10.00	False	False
’covariance type’: ’tied’, ’n ’: 11	0.79	0.79	0.32	0.10	-0.60	11.00	False	False
’covariance type’: ’diag’, ’n ’: 2	0.58	0.58	0.59	0.15	0.10	2.00	False	False
’covariance type’: ’diag’, ’n ’: 3	0.75	0.74	0.59	0.12	-0.44	3.00	False	False
’covariance type’: ’diag’, ’n ’: 4	0.88	0.82	0.49	0.15	0.10	4.00	False	False
’covariance type’: ’diag’, ’n ’: 5	0.97	0.97	0.64	0.21	0.06	5.00	True	True
’covariance type’: ’diag’, ’n ’: 6	0.93	0.94	0.57	0.13	-0.06	6.00	False	False
’covariance type’: ’diag’, ’n ’: 7	0.90	0.89	0.50	0.12	-0.31	7.00	False	False
’covariance type’: ’diag’, ’n ’: 8	0.87	0.87	0.43	0.10	-0.44	8.00	False	False
’covariance type’: ’diag’, ’n ’: 9	0.84	0.83	0.37	0.09	-0.52	9.00	False	False
’covariance type’: ’diag’, ’n ’: 10	0.81	0.81	0.32	0.08	-0.61	10.00	False	False
’covariance type’: ’diag’, ’n ’: 11	0.80	0.78	0.32	0.12	-0.53	11.00	False	False
’covariance type’: ’spherical’, ’n ’: 2	0.59	0.59	0.59	0.15	0.10	2.00	True	True
’covariance type’: ’spherical’, ’n ’: 3	0.73	0.69	0.46	0.14	-0.36	3.00	False	False
’covariance type’: ’spherical’, ’n ’: 4	0.82	0.82	0.50	0.15	0.08	4.00	False	False
’covariance type’: ’spherical’, ’n ’: 5	0.97	0.97	0.64	0.21	0.06	5.00	False	False

Table C.65: Synthetic dataset - GMM clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’covariance type’: ’full’, ’n ’: 11	0.78	0.78	0.36	0.10	-0.35	11.00	False	False
’covariance type’: ’tied’, ’n ’: 2	0.53	0.53	0.51	0.12	-0.81	2.00	False	False
’covariance type’: ’tied’, ’n ’: 3	0.74	0.74	0.49	0.12	0.47	3.00	False	False
’covariance type’: ’tied’, ’n ’: 4	0.87	0.88	0.59	0.15	-0.14	4.00	False	False
’covariance type’: ’tied’, ’n ’: 5	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’covariance type’: ’tied’, ’n ’: 6	0.93	0.93	0.67	0.16	0.53	6.00	False	False
’covariance type’: ’tied’, ’n ’: 7	0.89	0.89	0.58	0.13	0.30	7.00	False	False
’covariance type’: ’tied’, ’n ’: 8	0.87	0.86	0.53	0.11	0.03	8.00	False	False
’covariance type’: ’tied’, ’n ’: 9	0.83	0.83	0.52	0.12	-0.17	9.00	False	False
’covariance type’: ’tied’, ’n ’: 10	0.82	0.81	0.46	0.10	-0.16	10.00	False	False
’covariance type’: ’tied’, ’n ’: 11	0.78	0.78	0.36	0.09	-0.16	11.00	False	False
’covariance type’: ’diag’, ’n ’: 2	0.59	0.59	0.52	0.12	0.16	2.00	False	False
’covariance type’: ’diag’, ’n ’: 3	0.74	0.77	0.51	0.19	0.47	3.00	False	False
’covariance type’: ’diag’, ’n ’: 4	0.88	0.88	0.61	0.16	0.59	4.00	False	False
’covariance type’: ’diag’, ’n ’: 5	0.97	0.97	0.75	0.28	0.78	5.00	True	True
’covariance type’: ’diag’, ’n ’: 6	0.93	0.93	0.67	0.15	0.50	6.00	False	False
’covariance type’: ’diag’, ’n ’: 7	0.89	0.89	0.60	0.13	0.28	7.00	False	False

Table C.66: Synthetic dataset - GMM clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’covariance type’: ’full’, ’n ’: 2	0.59	0.59	0.73	0.21	0.55	2.00	False	False
’covariance type’: ’full’, ’n ’: 3	0.74	0.77	0.70	0.22	0.57	3.00	False	False
’covariance type’: ’full’, ’n ’: 4	0.87	0.88	0.75	0.29	0.72	4.00	False	False
’covariance type’: ’full’, ’n ’: 5	0.97	0.97	0.85	0.43	0.91	5.00	True	True
’covariance type’: ’full’, ’n ’: 6	0.93	0.93	0.76	0.17	0.68	6.00	False	False
’covariance type’: ’full’, ’n ’: 7	0.89	0.90	0.66	0.14	0.32	7.00	False	False
’covariance type’: ’full’, ’n ’: 8	0.86	0.86	0.57	0.13	0.42	8.00	False	False
’covariance type’: ’full’, ’n ’: 9	0.83	0.83	0.48	0.10	-0.26	9.00	False	False
’covariance type’: ’full’, ’n ’: 10	0.81	0.81	0.48	0.09	-0.46	10.00	False	False
’covariance type’: ’full’, ’n ’: 11	0.79	0.79	0.50	0.10	-0.39	11.00	False	False
’covariance type’: ’tied’, ’n ’: 2	0.59	0.59	0.73	0.21	0.55	2.00	False	False
’covariance type’: ’tied’, ’n ’: 3	0.77	0.77	0.70	0.22	0.70	3.00	False	False
’covariance type’: ’tied’, ’n ’: 4	0.88	0.88	0.75	0.29	0.72	4.00	False	False
’covariance type’: ’tied’, ’n ’: 5	0.97	0.97	0.85	0.43	0.91	5.00	False	False
’covariance type’: ’tied’, ’n ’: 6	0.93	0.93	0.75	0.18	0.62	6.00	False	False
’covariance type’: ’tied’, ’n ’: 7	0.89	0.89	0.66	0.16	0.37	7.00	False	False
’covariance type’: ’tied’, ’n ’: 8	0.86	0.87	0.57	0.12	0.05	8.00	False	False
’covariance type’: ’tied’, ’n ’: 9	0.83	0.83	0.46	0.10	-0.21	9.00	False	False
’covariance type’: ’tied’, ’n ’: 10	0.81	0.82	0.47	0.10	-0.54	10.00	False	False

Table C.67: Synthetic dataset - DBSCAN clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’eps’: 0.4, ’min samples’: 50	0.83	0.83	0.51	0.06	0.25	6.00	True	True
’eps’: 0.4, ’min samples’: 100	0.57	0.57	0.13	0.06	0.39	6.00	True	True
’eps’: 0.4, ’min samples’: 150	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 250	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 100	0.91	0.91	0.59	0.06	0.05	6.00	False	False
’eps’: 0.6, ’min samples’: 150	0.84	0.84	0.52	0.06	0.23	6.00	True	True
’eps’: 0.6, ’min samples’: 200	0.72	0.72	0.37	0.06	0.38	6.00	True	True
’eps’: 0.6, ’min samples’: 250	0.27	0.27	-0.09	0.12	0.16	3.00	True	True
’eps’: 0.6, ’min samples’: 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 350	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 200	0.92	0.92	0.60	0.06	0.05	6.00	True	True
’eps’: 0.8, ’min samples’: 250	0.90	0.90	0.58	0.06	0.13	6.00	True	True
’eps’: 0.8, ’min samples’: 300	0.83	0.84	0.52	0.06	0.19	6.00	True	True
’eps’: 0.8, ’min samples’: 350	0.67	0.67	0.29	0.08	0.37	5.00	True	True
’eps’: 0.8, ’min samples’: 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1, ’min samples’: 300	0.85	0.85	0.52	0.05	-0.34	5.00	False	False
’eps’: 1, ’min samples’: 350	0.92	0.92	0.60	0.06	0.04	6.00	True	True
’eps’: 1, ’min samples’: 400	0.90	0.90	0.58	0.06	0.12	6.00	True	True
’eps’: 1, ’min samples’: 450	0.80	0.80	0.48	0.06	0.29	6.00	True	True
’eps’: 1.2, ’min samples’: 100	0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.2, ’min samples’: 150	-0.00	0.00	0.17	0.16	nan	2.00	True	False
’eps’: 1.2, ’min samples’: 400	0.72	0.72	0.57	0.04	-0.06	4.00	False	False
’eps’: 1.2, ’min samples’: 450	0.91	0.91	0.60	0.06	-0.31	6.00	True	True
’eps’: 1.2, ’min samples’: 500	0.92	0.92	0.59	0.06	0.07	6.00	True	True
’eps’: 1.4, ’min samples’: 350	0.58	0.59	0.42	0.16	0.06	3.00	False	False
’eps’: 1.4, ’min samples’: 400	0.59	0.59	0.42	0.16	0.06	3.00	True	True
’eps’: 1.4, ’min samples’: 450	0.59	0.59	0.54	0.03	0.06	3.00	False	False

Table C.68: Synthetic dataset - DBSCAN clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’eps’: 0.2, ’min samples’: 650	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 50	0.83	0.83	0.51	0.06	0.25	6.00	True	True
’eps’: 0.4, ’min samples’: 100	0.57	0.57	0.13	0.06	0.39	6.00	True	True
’eps’: 0.4, ’min samples’: 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 350	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 450	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 550	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 650	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 50	0.71	0.71	0.55	0.06	-0.08	4.00	False	False
’eps’: 0.6, ’min samples’: 100	0.91	0.91	0.59	0.06	0.05	6.00	False	False
’eps’: 0.6, ’min samples’: 150	0.84	0.84	0.52	0.06	0.23	6.00	True	True
’eps’: 0.6, ’min samples’: 200	0.72	0.72	0.37	0.06	0.38	6.00	True	True
’eps’: 0.6, ’min samples’: 250	0.27	0.27	-0.09	0.12	0.16	3.00	True	True
’eps’: 0.6, ’min samples’: 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 350	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 450	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 550	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 650	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 50	0.59	0.59	0.54	0.03	0.06	3.00	False	False
’eps’: 0.8, ’min samples’: 100	0.58	0.58	0.54	0.03	0.06	3.00	False	False
’eps’: 0.8, ’min samples’: 150	0.85	0.85	0.52	0.06	-0.13	5.00	False	False
’eps’: 0.8, ’min samples’: 200	0.92	0.92	0.60	0.06	0.05	6.00	True	True
’eps’: 0.8, ’min samples’: 250	0.90	0.90	0.58	0.06	0.13	6.00	True	True
’eps’: 0.8, ’min samples’: 300	0.83	0.84	0.52	0.06	0.19	6.00	True	True
’eps’: 0.8, ’min samples’: 350	0.67	0.67	0.29	0.08	0.37	5.00	True	True
’eps’: 0.8, ’min samples’: 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1, ’min samples’: 50	-0.00	0.00	0.30	0.01	nan	2.00	False	False
’eps’: 1, ’min samples’: 100	0.58	0.58	0.56	0.03	0.06	3.00	False	False
’eps’: 1, ’min samples’: 150	0.59	0.59	0.54	0.03	0.06	3.00	False	False
’eps’: 1, ’min samples’: 200	0.59	0.59	0.55	0.03	0.06	3.00	False	False
’eps’: 1, ’min samples’: 250	0.71	0.71	0.55	0.05	-0.05	4.00	False	False
’eps’: 1, ’min samples’: 300	0.85	0.85	0.52	0.05	-0.34	5.00	False	False
’eps’: 1, ’min samples’: 350	0.92	0.92	0.60	0.06	0.04	6.00	True	True
’eps’: 1, ’min samples’: 400	0.90	0.90	0.58	0.06	0.12	6.00	True	True
’eps’: 1, ’min samples’: 450	0.80	0.80	0.48	0.06	0.29	6.00	True	True
’eps’: 1.2, ’min samples’: 100	0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.2, ’min samples’: 150	-0.00	0.00	0.17	0.16	nan	2.00	True	False
’eps’: 1.2, ’min samples’: 200	0.58	0.58	0.42	0.16	0.06	3.00	False	False
’eps’: 1.2, ’min samples’: 250	0.59	0.59	0.56	0.03	0.06	3.00	False	False
’eps’: 1.2, ’min samples’: 300	0.59	0.59	0.54	0.03	0.06	3.00	False	False
’eps’: 1.2, ’min samples’: 350	0.59	0.59	0.56	0.04	0.06	3.00	False	False
’eps’: 1.2, ’min samples’: 400	0.72	0.72	0.57	0.04	-0.06	4.00	False	False
’eps’: 1.2, ’min samples’: 450	0.91	0.91	0.60	0.06	-0.31	6.00	True	True
’eps’: 1.2, ’min samples’: 500	0.92	0.92	0.59	0.06	0.07	6.00	True	True
’eps’: 1.4, ’min samples’: 200	0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.4, ’min samples’: 250	0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.4, ’min samples’: 300	0.58	0.58	0.42	0.16	0.06	3.00	False	False
’eps’: 1.4, ’min samples’: 350	0.58	0.59	0.42	0.16	0.06	3.00	False	False
’eps’: 1.4, ’min samples’: 400	0.59	0.59	0.42	0.16	0.06	3.00	True	True
’eps’: 1.4, ’min samples’: 450	0.59	0.59	0.54	0.03	0.06	3.00	False	False

Table C.69: Synthetic dataset - DBSCAN clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’eps’: 0.2, ’min samples’: 650	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 50	0.83	0.83	0.51	0.06	0.25	6.00	True	True
’eps’: 0.4, ’min samples’: 100	0.57	0.57	0.13	0.06	0.39	6.00	True	True
’eps’: 0.4, ’min samples’: 150	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 250	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 650	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 50	0.71	0.71	0.55	0.06	-0.08	4.00	False	False
’eps’: 0.6, ’min samples’: 100	0.91	0.91	0.59	0.06	0.05	6.00	False	False
’eps’: 0.6, ’min samples’: 150	0.84	0.84	0.52	0.06	0.23	6.00	True	True
’eps’: 0.6, ’min samples’: 200	0.72	0.72	0.37	0.06	0.38	6.00	True	True
’eps’: 0.6, ’min samples’: 250	0.27	0.27	-0.09	0.12	0.16	3.00	True	True
’eps’: 0.6, ’min samples’: 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 350	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 450	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 550	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 650	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 50	0.59	0.59	0.54	0.03	0.06	3.00	False	False
’eps’: 0.8, ’min samples’: 100	0.58	0.58	0.52	0.02	0.06	3.00	False	False
’eps’: 0.8, ’min samples’: 150	0.85	0.85	0.52	0.06	-0.13	5.00	False	False
’eps’: 0.8, ’min samples’: 200	0.92	0.92	0.60	0.06	0.05	6.00	True	True
’eps’: 0.8, ’min samples’: 250	0.90	0.90	0.58	0.06	0.13	6.00	True	True
’eps’: 0.8, ’min samples’: 300	0.83	0.84	0.52	0.06	0.19	6.00	True	True
’eps’: 0.8, ’min samples’: 350	0.67	0.67	0.29	0.08	0.37	5.00	True	True
’eps’: 0.8, ’min samples’: 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 450	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 550	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 650	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1, ’min samples’: 50	-0.00	0.00	0.30	0.01	nan	2.00	False	False
’eps’: 1, ’min samples’: 100	0.58	0.58	0.56	0.03	0.06	3.00	False	False
’eps’: 1, ’min samples’: 150	0.59	0.59	0.54	0.03	0.06	3.00	False	False
’eps’: 1, ’min samples’: 200	0.59	0.59	0.55	0.03	0.06	3.00	False	False
’eps’: 1, ’min samples’: 250	0.71	0.71	0.56	0.05	-0.05	4.00	False	False
’eps’: 1, ’min samples’: 300	0.85	0.85	0.52	0.05	-0.34	5.00	False	False
’eps’: 1, ’min samples’: 350	0.92	0.92	0.60	0.06	0.04	6.00	True	True
’eps’: 1, ’min samples’: 400	0.90	0.90	0.58	0.06	0.12	6.00	True	True
’eps’: 1, ’min samples’: 450	0.80	0.80	0.48	0.06	0.29	6.00	True	True
’eps’: 1.2, ’min samples’: 100	0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.2, ’min samples’: 150	-0.00	0.00	0.17	0.16	nan	2.00	True	False
’eps’: 1.2, ’min samples’: 200	0.58	0.58	0.42	0.16	0.06	3.00	False	False
’eps’: 1.2, ’min samples’: 250	0.59	0.59	0.56	0.03	0.06	3.00	False	False
’eps’: 1.2, ’min samples’: 300	0.59	0.59	0.54	0.03	0.06	3.00	False	False
’eps’: 1.2, ’min samples’: 350	0.59	0.59	0.56	0.04	0.06	3.00	False	False
’eps’: 1.2, ’min samples’: 400	0.72	0.72	0.57	0.04	-0.06	4.00	False	False
’eps’: 1.2, ’min samples’: 450	0.91	0.91	0.60	0.06	-0.31	6.00	False	False
’eps’: 1.2, ’min samples’: 500	0.92	0.92	0.59	0.06	0.07	6.00	True	True
’eps’: 1.4, ’min samples’: 250	0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.4, ’min samples’: 300	0.58	0.58	0.42	0.16	0.06	3.00	False	False
’eps’: 1.4, ’min samples’: 350	0.58	0.59	0.42	0.16	0.06	3.00	False	False
’eps’: 1.4, ’min samples’: 400	0.59	0.59	0.42	0.16	0.06	3.00	True	True
’eps’: 1.4, ’min samples’: 450	0.59	0.59	0.54	0.03	0.06	3.00	False	False

Table C.70: Synthetic dataset - DBSCAN clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’eps’: 0.2, ’min samples’: 650	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 50	0.86	0.86	0.56	0.06	0.45	6.00	True	True
’eps’: 0.4, ’min samples’: 100	0.63	0.63	0.24	0.06	0.46	6.00	True	True
’eps’: 0.4, ’min samples’: 150	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.4, ’min samples’: 650	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 50	0.72	0.72	0.57	0.06	0.04	4.00	False	False
’eps’: 0.6, ’min samples’: 100	0.92	0.92	0.62	0.05	0.20	6.00	True	True
’eps’: 0.6, ’min samples’: 150	0.87	0.87	0.57	0.05	0.37	6.00	True	True
’eps’: 0.6, ’min samples’: 200	0.76	0.76	0.44	0.06	0.50	6.00	True	True
’eps’: 0.6, ’min samples’: 250	0.47	0.48	0.04	0.06	0.28	4.00	False	False
’eps’: 0.6, ’min samples’: 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.6, ’min samples’: 650	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 50	0.59	0.59	0.56	0.03	0.10	3.00	False	False
’eps’: 0.8, ’min samples’: 100	0.72	0.72	0.59	0.04	0.01	4.00	False	False
’eps’: 0.8, ’min samples’: 150	0.86	0.86	0.54	0.05	-0.25	5.00	False	False
’eps’: 0.8, ’min samples’: 200	0.94	0.94	0.63	0.05	0.13	6.00	True	True
’eps’: 0.8, ’min samples’: 250	0.92	0.92	0.62	0.05	0.19	6.00	False	False
’eps’: 0.8, ’min samples’: 300	0.87	0.87	0.58	0.05	0.37	6.00	True	True
’eps’: 0.8, ’min samples’: 350	0.76	0.76	0.44	0.05	0.50	6.00	False	False
’eps’: 0.8, ’min samples’: 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 450	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 550	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 650	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1, ’min samples’: 50	-0.00	0.00	0.15	0.16	nan	2.00	False	False
’eps’: 1, ’min samples’: 100	0.59	0.59	0.41	0.16	0.10	3.00	True	True
’eps’: 1, ’min samples’: 150	0.59	0.59	0.41	0.16	0.10	3.00	True	True
’eps’: 1, ’min samples’: 200	0.72	0.72	0.59	0.03	0.08	4.00	False	False
’eps’: 1, ’min samples’: 250	0.72	0.72	0.57	0.05	0.12	4.00	False	False
’eps’: 1, ’min samples’: 300	0.94	0.94	0.63	0.06	-0.19	6.00	True	True
’eps’: 1, ’min samples’: 350	0.94	0.94	0.63	0.05	0.13	6.00	True	True
’eps’: 1, ’min samples’: 400	0.92	0.92	0.62	0.04	0.13	6.00	False	False
’eps’: 1, ’min samples’: 450	0.85	0.85	0.54	0.05	0.28	6.00	False	False
’eps’: 1, ’min samples’: 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1, ’min samples’: 550	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1, ’min samples’: 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1, ’min samples’: 650	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1.2, ’min samples’: 50	0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.2, ’min samples’: 100	0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.2, ’min samples’: 150	0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.2, ’min samples’: 200	0.58	0.59	0.41	0.16	0.10	3.00	False	False
’eps’: 1.2, ’min samples’: 250	0.59	0.59	0.41	0.16	0.10	3.00	False	False
’eps’: 1.2, ’min samples’: 300	0.59	0.59	0.41	0.16	0.10	3.00	False	False
’eps’: 1.2, ’min samples’: 350	0.72	0.72	0.59	0.04	0.04	4.00	False	False
’eps’: 1.2, ’min samples’: 400	0.73	0.73	0.58	0.05	0.01	4.00	False	False
’eps’: 1.2, ’min samples’: 450	0.93	0.93	0.63	0.06	-0.22	6.00	True	True
’eps’: 1.4, ’min samples’: 150	0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.4, ’min samples’: 200	0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.4, ’min samples’: 250	0.00	0.00	0.00	0.00	nan	1.00	False	False
’eps’: 1.4, ’min samples’: 300	0.58	0.58	0.59	0.15	0.10	2.00	True	True
’eps’: 1.4, ’min samples’: 350	0.58	0.59	0.41	0.16	0.10	3.00	False	False
’eps’: 1.4, ’min samples’: 400	0.59	0.59	0.41	0.16	0.10	3.00	False	False

Table C.71: Synthetic dataset - DBSCAN clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'eps': 4, 'min samples': 50	0.97	0.97	0.75	0.28	0.78	5.00	True	True
'eps': 4, 'min samples': 100	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 4, 'min samples': 150	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 4, 'min samples': 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 4, 'min samples': 250	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 4, 'min samples': 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 4, 'min samples': 350	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 4, 'min samples': 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 4, 'min samples': 450	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 5, 'min samples': 50	0.97	0.97	0.75	0.28	0.78	5.00	False	False
'eps': 5, 'min samples': 100	0.96	0.96	0.74	0.07	0.78	6.00	False	False
'eps': 5, 'min samples': 150	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 5, 'min samples': 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 5, 'min samples': 250	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 5, 'min samples': 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 5, 'min samples': 350	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 5, 'min samples': 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 5, 'min samples': 450	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 6, 'min samples': 50	0.97	0.97	0.75	0.28	0.78	5.00	False	False
'eps': 6, 'min samples': 100	0.97	0.97	0.75	0.28	0.78	5.00	False	False
'eps': 6, 'min samples': 150	0.94	0.94	0.71	0.07	0.77	6.00	False	False
'eps': 6, 'min samples': 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 6, 'min samples': 250	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 6, 'min samples': 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 6, 'min samples': 350	0.00	0.00	0.00	0.00	0.00	1.00	False	False

Table C.72: Synthetic dataset - DBSCAN clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'eps': 0.5, 'min samples': 50	0.97	0.97	0.85	0.43	0.91	5.00	True	True
'eps': 0.5, 'min samples': 100	0.97	0.97	0.85	0.43	0.91	5.00	False	False
'eps': 0.5, 'min samples': 150	0.92	0.92	0.73	0.20	0.65	7.00	False	False
'eps': 0.5, 'min samples': 200	0.39	0.39	-0.14	0.10	0.15	5.00	False	False
'eps': 0.5, 'min samples': 250	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.5, 'min samples': 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.5, 'min samples': 350	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.5, 'min samples': 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 0.5, 'min samples': 450	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 1, 'min samples': 50	0.97	0.97	0.85	0.43	0.91	5.00	False	False
'eps': 1, 'min samples': 100	0.97	0.97	0.85	0.43	0.91	5.00	False	False
'eps': 1, 'min samples': 150	0.97	0.97	0.85	0.43	0.91	5.00	False	False
'eps': 1, 'min samples': 200	0.97	0.97	0.85	0.43	0.91	5.00	False	False
'eps': 1, 'min samples': 250	0.97	0.97	0.85	0.43	0.91	5.00	False	False
'eps': 1, 'min samples': 300	0.97	0.97	0.85	0.43	0.91	5.00	False	False
'eps': 1, 'min samples': 350	0.97	0.97	0.85	0.43	0.91	5.00	False	False
'eps': 1, 'min samples': 400	0.97	0.97	0.85	0.43	0.91	5.00	False	False
'eps': 1, 'min samples': 450	0.97	0.97	0.85	0.43	0.91	5.00	False	False

Table C.73: Synthetic dataset - OPTICS clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'max eps': 0.5, 'min samples': 50	0.84	0.84	0.51	0.05	-0.08	5.00	False	False
'max eps': 0.5, 'min samples': 100	0.82	0.82	0.50	0.06	0.42	6.00	True	True
'max eps': 0.5, 'min samples': 150	0.63	0.63	0.24	0.06	0.43	6.00	False	False
'max eps': 0.5, 'min samples': 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 0.5, 'min samples': 250	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 0.5, 'min samples': 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 0.5, 'min samples': 350	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 0.5, 'min samples': 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 0.5, 'min samples': 450	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 0.5, 'min samples': 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 0.7, 'min samples': 50	0.72	0.72	0.57	0.19	0.29	3.00	False	False
'max eps': 0.7, 'min samples': 100	0.71	0.71	0.55	0.06	0.02	4.00	False	False
'max eps': 0.7, 'min samples': 150	0.91	0.91	0.59	0.06	0.05	6.00	True	True
'max eps': 0.7, 'min samples': 200	0.87	0.87	0.55	0.06	0.09	6.00	False	False
'max eps': 0.7, 'min samples': 250	0.78	0.78	0.46	0.06	0.41	6.00	False	False
'max eps': 0.7, 'min samples': 300	0.53	0.53	0.08	0.08	0.32	5.00	False	False
'max eps': 0.7, 'min samples': 350	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 0.7, 'min samples': 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 0.7, 'min samples': 450	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 0.7, 'min samples': 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1, 'min samples': 50	0.72	0.72	0.57	0.20	0.29	3.00	False	False
'max eps': 1, 'min samples': 100	0.71	0.71	0.57	0.20	0.30	3.00	True	True
'max eps': 1, 'min samples': 150	0.59	0.59	0.54	0.03	0.06	3.00	False	False
'max eps': 1, 'min samples': 200	0.59	0.59	0.55	0.03	0.06	3.00	False	False
'max eps': 1, 'min samples': 250	0.71	0.71	0.55	0.05	-0.06	4.00	False	False
'max eps': 1, 'min samples': 300	0.85	0.85	0.52	0.06	-0.32	5.00	False	False
'max eps': 1, 'min samples': 350	0.92	0.92	0.60	0.06	0.02	6.00	True	True
'max eps': 1, 'min samples': 400	0.90	0.90	0.58	0.06	0.04	6.00	False	False
'max eps': 1, 'min samples': 450	0.78	0.78	0.45	0.06	0.33	6.00	False	False
'max eps': 1, 'min samples': 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 1.2, 'min samples': 50	0.72	0.72	0.57	0.20	0.29	3.00	False	False
'max eps': 1.2, 'min samples': 100	0.72	0.72	0.57	0.20	0.30	3.00	False	False
'max eps': 1.2, 'min samples': 150	0.58	0.58	0.48	0.02	0.06	3.00	False	False
'max eps': 1.2, 'min samples': 200	0.58	0.58	0.42	0.16	0.06	3.00	False	False
'max eps': 1.2, 'min samples': 250	0.59	0.59	0.56	0.03	0.06	3.00	False	False
'max eps': 1.2, 'min samples': 300	0.59	0.59	0.54	0.03	0.06	3.00	False	False
'max eps': 1.2, 'min samples': 350	0.73	0.73	0.48	0.13	0.48	3.00	True	True
'max eps': 1.2, 'min samples': 400	0.82	0.82	0.49	0.14	0.05	4.00	False	False
'max eps': 1.2, 'min samples': 450	0.91	0.91	0.60	0.06	-0.34	6.00	False	False
'max eps': 1.2, 'min samples': 500	0.90	0.90	0.58	0.06	0.12	6.00	True	True

Table C.74: Synthetic dataset - OPTICS clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’max eps’: 0.5, ’min samples’: 50	0.84	0.84	0.51	0.05	-0.08	5.00	False	False
’max eps’: 0.5, ’min samples’: 100	0.82	0.82	0.50	0.06	0.42	6.00	True	True
’max eps’: 0.5, ’min samples’: 150	0.63	0.63	0.24	0.06	0.43	6.00	False	False
’max eps’: 0.5, ’min samples’: 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 250	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 350	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 450	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.7, ’min samples’: 50	0.72	0.72	0.57	0.19	0.29	3.00	False	False
’max eps’: 0.7, ’min samples’: 100	0.71	0.71	0.55	0.06	0.02	4.00	False	False
’max eps’: 0.7, ’min samples’: 150	0.91	0.91	0.59	0.06	0.05	6.00	True	True
’max eps’: 0.7, ’min samples’: 200	0.87	0.87	0.55	0.06	0.09	6.00	False	False
’max eps’: 0.7, ’min samples’: 250	0.78	0.78	0.46	0.06	0.41	6.00	False	False
’max eps’: 0.7, ’min samples’: 300	0.53	0.53	0.08	0.08	0.32	5.00	False	False
’max eps’: 0.7, ’min samples’: 350	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.7, ’min samples’: 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.7, ’min samples’: 450	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.7, ’min samples’: 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 1, ’min samples’: 50	0.72	0.72	0.57	0.20	0.29	3.00	False	False
’max eps’: 1, ’min samples’: 100	0.71	0.71	0.57	0.20	0.30	3.00	True	True
’max eps’: 1, ’min samples’: 150	0.59	0.59	0.54	0.03	0.06	3.00	False	False
’max eps’: 1, ’min samples’: 200	0.59	0.59	0.55	0.03	0.06	3.00	False	False
’max eps’: 1, ’min samples’: 250	0.71	0.71	0.55	0.05	-0.06	4.00	False	False
’max eps’: 1, ’min samples’: 300	0.85	0.85	0.52	0.06	-0.32	5.00	False	False
’max eps’: 1, ’min samples’: 350	0.92	0.92	0.60	0.06	0.02	6.00	True	True
’max eps’: 1, ’min samples’: 400	0.90	0.90	0.58	0.06	0.04	6.00	False	False
’max eps’: 1, ’min samples’: 450	0.78	0.78	0.45	0.06	0.33	6.00	False	False
’max eps’: 1, ’min samples’: 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 1.2, ’min samples’: 50	0.72	0.72	0.57	0.20	0.29	3.00	False	False
’max eps’: 1.2, ’min samples’: 100	0.72	0.72	0.57	0.20	0.30	3.00	False	False
’max eps’: 1.2, ’min samples’: 150	0.58	0.58	0.48	0.02	0.06	3.00	False	False
’max eps’: 1.2, ’min samples’: 200	0.58	0.58	0.42	0.16	0.06	3.00	False	False
’max eps’: 1.2, ’min samples’: 250	0.59	0.59	0.56	0.03	0.06	3.00	False	False
’max eps’: 1.2, ’min samples’: 300	0.59	0.59	0.54	0.03	0.06	3.00	False	False
’max eps’: 1.2, ’min samples’: 350	0.73	0.73	0.48	0.13	0.48	3.00	True	True
’max eps’: 1.2, ’min samples’: 400	0.82	0.82	0.49	0.14	0.05	4.00	False	False
’max eps’: 1.2, ’min samples’: 450	0.91	0.91	0.60	0.06	-0.34	6.00	False	False
’max eps’: 1.2, ’min samples’: 500	0.90	0.90	0.58	0.06	0.12	6.00	True	True

Table C.75: Synthetic dataset - OPTICS clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’max eps’: 0.5, ’min samples’: 50	0.84	0.84	0.51	0.06	-0.08	5.00	False	False
’max eps’: 0.5, ’min samples’: 100	0.82	0.82	0.50	0.06	0.42	6.00	True	True
’max eps’: 0.5, ’min samples’: 150	0.63	0.63	0.24	0.06	0.43	6.00	False	False
’max eps’: 0.5, ’min samples’: 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 250	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 350	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 450	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.7, ’min samples’: 50	0.69	0.69	0.29	0.12	0.34	3.00	False	False
’max eps’: 0.7, ’min samples’: 100	0.71	0.71	0.55	0.06	0.02	4.00	False	False
’max eps’: 0.7, ’min samples’: 150	0.91	0.91	0.59	0.06	0.05	6.00	True	True
’max eps’: 0.7, ’min samples’: 200	0.87	0.87	0.55	0.06	0.09	6.00	False	False
’max eps’: 0.7, ’min samples’: 250	0.78	0.78	0.46	0.06	0.41	6.00	False	False
’max eps’: 0.7, ’min samples’: 300	0.53	0.53	0.08	0.08	0.32	5.00	False	False
’max eps’: 0.7, ’min samples’: 350	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.7, ’min samples’: 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.7, ’min samples’: 450	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.7, ’min samples’: 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 1, ’min samples’: 50	0.69	0.69	0.29	0.12	0.34	3.00	False	False
’max eps’: 1, ’min samples’: 100	0.71	0.71	0.57	0.20	0.30	3.00	True	True
’max eps’: 1, ’min samples’: 150	0.59	0.59	0.54	0.03	0.06	3.00	False	False
’max eps’: 1, ’min samples’: 200	0.59	0.59	0.55	0.03	0.06	3.00	False	False
’max eps’: 1, ’min samples’: 250	0.71	0.71	0.56	0.05	-0.06	4.00	False	False
’max eps’: 1, ’min samples’: 300	0.85	0.85	0.52	0.06	-0.32	5.00	False	False
’max eps’: 1, ’min samples’: 350	0.92	0.92	0.60	0.06	0.02	6.00	True	True
’max eps’: 1, ’min samples’: 400	0.90	0.90	0.58	0.06	0.04	6.00	False	False
’max eps’: 1, ’min samples’: 450	0.78	0.78	0.45	0.06	0.33	6.00	False	False
’max eps’: 1, ’min samples’: 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 1.2, ’min samples’: 50	0.69	0.69	0.29	0.12	0.34	3.00	False	False
’max eps’: 1.2, ’min samples’: 100	0.72	0.72	0.57	0.20	0.30	3.00	False	False
’max eps’: 1.2, ’min samples’: 150	0.58	0.58	0.48	0.02	0.06	3.00	False	False
’max eps’: 1.2, ’min samples’: 200	0.58	0.58	0.42	0.16	0.06	3.00	False	False
’max eps’: 1.2, ’min samples’: 250	0.59	0.59	0.56	0.03	0.06	3.00	False	False
’max eps’: 1.2, ’min samples’: 300	0.59	0.59	0.54	0.03	0.06	3.00	False	False
’max eps’: 1.2, ’min samples’: 350	0.74	0.74	0.49	0.13	0.48	3.00	True	True
’max eps’: 1.2, ’min samples’: 400	0.82	0.82	0.49	0.14	0.05	4.00	False	False
’max eps’: 1.2, ’min samples’: 450	0.91	0.91	0.60	0.06	-0.34	6.00	False	False
’max eps’: 1.2, ’min samples’: 500	0.90	0.90	0.58	0.06	0.12	6.00	True	True

Table C.76: Synthetic dataset - OPTICS clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’max eps’: 0.5, ’min samples’: 50	0.93	0.93	0.62	0.06	0.13	6.00	True	True
’max eps’: 0.5, ’min samples’: 100	0.85	0.85	0.55	0.06	0.37	6.00	False	False
’max eps’: 0.5, ’min samples’: 150	0.68	0.68	0.33	0.06	0.49	6.00	True	True
’max eps’: 0.5, ’min samples’: 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 350	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 450	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.7, ’min samples’: 50	0.72	0.73	0.59	0.21	0.44	3.00	False	False
’max eps’: 0.7, ’min samples’: 100	0.85	0.85	0.53	0.15	0.16	4.00	False	False
’max eps’: 0.7, ’min samples’: 150	0.93	0.93	0.62	0.05	0.13	6.00	True	True
’max eps’: 0.7, ’min samples’: 200	0.89	0.89	0.59	0.05	0.16	6.00	True	True
’max eps’: 1, ’min samples’: 50	0.73	0.73	0.59	0.21	0.44	3.00	True	True
’max eps’: 1, ’min samples’: 100	0.68	0.68	0.26	0.10	0.26	3.00	False	False
’max eps’: 1, ’min samples’: 150	0.85	0.85	0.53	0.15	0.23	4.00	False	False
’max eps’: 1, ’min samples’: 200	0.72	0.72	0.59	0.03	0.08	4.00	False	False
’max eps’: 1, ’min samples’: 250	0.85	0.85	0.53	0.16	0.12	4.00	False	False
’max eps’: 1, ’min samples’: 300	0.94	0.94	0.63	0.06	-0.12	6.00	True	True
’max eps’: 1, ’min samples’: 350	0.94	0.94	0.63	0.05	0.11	6.00	True	True
’max eps’: 1.2, ’min samples’: 100	0.68	0.68	0.26	0.10	0.26	3.00	False	False
’max eps’: 1.2, ’min samples’: 150	0.84	0.84	0.53	0.15	0.23	4.00	False	False
’max eps’: 1.2, ’min samples’: 200	0.71	0.71	0.53	0.10	0.11	4.00	False	False
’max eps’: 1.2, ’min samples’: 250	0.67	0.67	0.25	0.10	0.26	3.00	False	False
’max eps’: 1.5, ’min samples’: 300	0.65	0.65	0.25	0.10	0.26	3.00	False	False
’max eps’: 1.5, ’min samples’: 350	0.70	0.70	0.59	0.21	0.45	3.00	True	True
’max eps’: 1.5, ’min samples’: 400	0.63	0.63	0.24	0.10	0.26	3.00	False	False
’max eps’: 1.5, ’min samples’: 450	0.62	0.62	0.24	0.10	0.26	3.00	False	False
’max eps’: 1.5, ’min samples’: 500	0.70	0.70	0.58	0.21	0.46	3.00	True	True

Table C.77: Synthetic dataset - OPTICS clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’max eps’: 4, ’min samples’: 50	0.96	0.96	0.73	0.07	0.77	6.00	False	False
’max eps’: 4, ’min samples’: 100	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 4, ’min samples’: 150	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 4, ’min samples’: 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 4.5, ’min samples’: 50	0.97	0.97	0.75	0.28	0.78	5.00	True	True
’max eps’: 4.5, ’min samples’: 100	0.27	0.27	-0.29	0.07	0.18	5.00	False	False
’max eps’: 4.5, ’min samples’: 150	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 4.5, ’min samples’: 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 5, ’min samples’: 50	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’max eps’: 5, ’min samples’: 100	0.95	0.95	0.74	0.06	0.77	6.00	False	False
’max eps’: 5, ’min samples’: 150	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 5, ’min samples’: 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 5.5, ’min samples’: 50	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’max eps’: 5.5, ’min samples’: 100	0.96	0.96	0.73	0.07	0.77	6.00	False	False
’max eps’: 5.5, ’min samples’: 150	0.10	0.10	-0.14	0.04	nan	2.00	False	False
’max eps’: 5.5, ’min samples’: 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 6, ’min samples’: 50	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’max eps’: 6, ’min samples’: 100	0.96	0.96	0.73	0.07	0.77	6.00	False	False
’max eps’: 6, ’min samples’: 150	0.93	0.93	0.70	0.07	0.75	6.00	False	False
’max eps’: 6, ’min samples’: 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False

Table C.78: Synthetic dataset - OPTICS clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’max eps’: 0.5, ’min samples’: 50	0.33	0.34	-0.43	0.09	0.13	13.00	False	False
’max eps’: 0.5, ’min samples’: 100	0.93	0.93	0.76	0.22	0.80	6.00	False	False
’max eps’: 0.5, ’min samples’: 150	0.88	0.88	0.61	0.14	0.53	7.00	False	False
’max eps’: 0.5, ’min samples’: 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 250	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 350	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 0.5, ’min samples’: 450	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 1, ’min samples’: 50	0.33	0.34	-0.43	0.09	0.13	13.00	False	False
’max eps’: 1, ’min samples’: 100	0.93	0.93	0.76	0.22	0.80	6.00	False	False
’max eps’: 1, ’min samples’: 150	0.84	0.84	0.45	0.13	0.68	6.00	False	False
’max eps’: 1, ’min samples’: 200	0.97	0.97	0.85	0.43	0.91	5.00	True	True
’max eps’: 1, ’min samples’: 250	0.97	0.97	0.85	0.43	0.91	5.00	False	False
’max eps’: 1, ’min samples’: 300	0.96	0.96	0.70	0.04	0.90	6.00	False	False
’max eps’: 1, ’min samples’: 350	0.96	0.96	0.84	0.09	0.90	6.00	False	False
’max eps’: 1, ’min samples’: 400	0.96	0.96	0.84	0.08	0.90	6.00	False	False
’max eps’: 1, ’min samples’: 450	0.95	0.95	0.83	0.09	0.89	6.00	False	False

Table C.79: Synthetic dataset - HDBSCAN clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’min cluster size’: 5, ’min samples’: 5	0.74	0.74	0.48	0.03	0.09	4.00	False	False
’min cluster size’: 5, ’min samples’: 7	0.82	0.82	0.49	0.07	-0.04	5.00	False	False
’min cluster size’: 5, ’min samples’: 10	0.92	0.92	0.59	0.05	0.13	6.00	True	True
’min cluster size’: 5, ’min samples’: 20	0.87	0.87	0.56	0.06	0.13	6.00	True	True
’min cluster size’: 5, ’min samples’: 30	0.86	0.86	0.54	0.07	0.24	6.00	True	True
’min cluster size’: 5, ’min samples’: 40	0.85	0.85	0.54	0.06	0.19	6.00	False	False
’min cluster size’: 10, ’min samples’: 5	0.86	0.86	0.45	0.06	0.12	7.00	False	False
’min cluster size’: 10, ’min samples’: 7	0.82	0.82	0.49	0.07	-0.04	5.00	False	False
’min cluster size’: 10, ’min samples’: 10	0.92	0.92	0.59	0.05	0.13	6.00	False	False
’min cluster size’: 10, ’min samples’: 20	0.87	0.87	0.56	0.06	0.13	6.00	False	False
’min cluster size’: 10, ’min samples’: 30	0.86	0.86	0.54	0.07	0.24	6.00	False	False
’min cluster size’: 10, ’min samples’: 40	0.85	0.85	0.54	0.06	0.19	6.00	False	False
’min cluster size’: 15, ’min samples’: 5	0.87	0.87	0.55	0.06	0.19	6.00	True	True
’min cluster size’: 15, ’min samples’: 7	0.86	0.86	0.55	0.07	0.19	6.00	True	True
’min cluster size’: 15, ’min samples’: 10	0.92	0.92	0.59	0.05	0.13	6.00	False	False
’min cluster size’: 15, ’min samples’: 20	0.87	0.87	0.56	0.06	0.13	6.00	False	False
’min cluster size’: 15, ’min samples’: 30	0.86	0.86	0.54	0.07	0.24	6.00	False	False
’min cluster size’: 15, ’min samples’: 40	0.85	0.85	0.54	0.06	0.19	6.00	False	False
’min cluster size’: 20, ’min samples’: 5	0.87	0.87	0.55	0.06	0.19	6.00	False	False
’min cluster size’: 20, ’min samples’: 7	0.86	0.86	0.55	0.07	0.19	6.00	False	False
’min cluster size’: 20, ’min samples’: 10	0.92	0.92	0.59	0.05	0.13	6.00	False	False
’min cluster size’: 20, ’min samples’: 20	0.87	0.87	0.56	0.06	0.13	6.00	False	False
’min cluster size’: 20, ’min samples’: 30	0.86	0.86	0.54	0.07	0.24	6.00	False	False
’min cluster size’: 20, ’min samples’: 40	0.85	0.85	0.54	0.06	0.19	6.00	False	False

Table C.80: Synthetic dataset - HDBSCAN clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'min cluster size': 5, 'min samples': 5	0.74	0.74	0.48	0.03	0.09	4.00	False	False
'min cluster size': 5, 'min samples': 7	0.73	0.73	0.57	0.05	-0.08	4.00	False	False
'min cluster size': 5, 'min samples': 10	0.91	0.91	0.59	0.06	0.13	6.00	True	True
'min cluster size': 5, 'min samples': 20	0.92	0.92	0.60	0.05	0.12	6.00	True	True
'min cluster size': 5, 'min samples': 30	0.85	0.85	0.54	0.06	0.19	6.00	False	False
'min cluster size': 5, 'min samples': 40	0.87	0.87	0.55	0.07	0.19	6.00	True	True
'min cluster size': 10, 'min samples': 5	0.86	0.86	0.45	0.07	0.12	7.00	False	False
'min cluster size': 10, 'min samples': 7	0.89	0.89	0.47	0.07	0.09	7.00	False	False
'min cluster size': 10, 'min samples': 10	0.91	0.91	0.59	0.06	0.13	6.00	False	False
'min cluster size': 10, 'min samples': 20	0.92	0.92	0.60	0.05	0.12	6.00	False	False
'min cluster size': 10, 'min samples': 30	0.85	0.85	0.54	0.06	0.19	6.00	False	False
'min cluster size': 10, 'min samples': 40	0.87	0.87	0.55	0.07	0.19	6.00	False	False
'min cluster size': 15, 'min samples': 5	0.87	0.87	0.56	0.06	0.19	6.00	True	True
'min cluster size': 15, 'min samples': 7	0.90	0.90	0.57	0.07	0.13	6.00	True	True
'min cluster size': 15, 'min samples': 10	0.91	0.91	0.59	0.06	0.13	6.00	False	False
'min cluster size': 15, 'min samples': 20	0.92	0.92	0.60	0.05	0.12	6.00	False	False
'min cluster size': 15, 'min samples': 30	0.85	0.85	0.54	0.06	0.19	6.00	False	False
'min cluster size': 15, 'min samples': 40	0.87	0.87	0.55	0.07	0.19	6.00	False	False
'min cluster size': 20, 'min samples': 5	0.87	0.87	0.56	0.06	0.19	6.00	False	False
'min cluster size': 20, 'min samples': 7	0.90	0.90	0.57	0.07	0.13	6.00	False	False
'min cluster size': 20, 'min samples': 10	0.91	0.91	0.59	0.06	0.13	6.00	False	False
'min cluster size': 20, 'min samples': 20	0.92	0.92	0.60	0.05	0.12	6.00	False	False
'min cluster size': 20, 'min samples': 30	0.85	0.85	0.54	0.06	0.19	6.00	False	False
'min cluster size': 20, 'min samples': 40	0.87	0.87	0.55	0.07	0.19	6.00	False	False

Table C.81: Synthetic dataset - HDBSCAN clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'min cluster size': 5, 'min samples': 5	0.83	0.83	0.54	0.06	0.32	5.00	True	True
'min cluster size': 5, 'min samples': 7	0.87	0.87	0.56	0.06	0.19	6.00	True	True
'min cluster size': 5, 'min samples': 10	0.69	0.69	0.54	0.06	-0.02	4.00	False	False
'min cluster size': 5, 'min samples': 20	0.90	0.90	0.58	0.05	0.12	6.00	True	True
'min cluster size': 5, 'min samples': 30	0.85	0.85	0.53	0.06	0.19	6.00	False	False
'min cluster size': 5, 'min samples': 40	0.83	0.83	0.51	0.07	0.22	6.00	True	True
'min cluster size': 10, 'min samples': 5	0.87	0.87	0.45	0.06	0.11	7.00	False	False
'min cluster size': 10, 'min samples': 7	0.87	0.87	0.56	0.06	0.19	6.00	False	False
'min cluster size': 10, 'min samples': 10	0.86	0.86	0.55	0.06	0.20	6.00	True	True
'min cluster size': 10, 'min samples': 20	0.90	0.90	0.58	0.05	0.12	6.00	False	False
'min cluster size': 10, 'min samples': 30	0.85	0.85	0.53	0.06	0.19	6.00	False	False
'min cluster size': 10, 'min samples': 40	0.83	0.83	0.51	0.07	0.22	6.00	False	False
'min cluster size': 15, 'min samples': 5	0.89	0.89	0.58	0.06	0.14	6.00	True	True
'min cluster size': 15, 'min samples': 7	0.87	0.87	0.56	0.06	0.19	6.00	False	False
'min cluster size': 15, 'min samples': 10	0.86	0.86	0.55	0.06	0.20	6.00	False	False
'min cluster size': 15, 'min samples': 20	0.90	0.90	0.58	0.05	0.12	6.00	False	False
'min cluster size': 15, 'min samples': 30	0.85	0.85	0.53	0.06	0.19	6.00	False	False
'min cluster size': 15, 'min samples': 40	0.83	0.83	0.51	0.07	0.22	6.00	False	False
'min cluster size': 20, 'min samples': 5	0.89	0.89	0.58	0.06	0.14	6.00	False	False
'min cluster size': 20, 'min samples': 7	0.87	0.87	0.56	0.06	0.19	6.00	False	False
'min cluster size': 20, 'min samples': 10	0.86	0.86	0.55	0.06	0.20	6.00	False	False
'min cluster size': 20, 'min samples': 20	0.90	0.90	0.58	0.05	0.12	6.00	False	False
'min cluster size': 20, 'min samples': 30	0.85	0.85	0.53	0.06	0.19	6.00	False	False
'min cluster size': 20, 'min samples': 40	0.83	0.83	0.51	0.07	0.22	6.00	False	False

Table C.82: Synthetic dataset - HDBSCAN clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’min cluster size’: 5, ’min samples’: 5	0.73	0.73	0.57	0.06	0.05	4.00	False	False
’min cluster size’: 5, ’min samples’: 7	0.73	0.73	0.57	0.06	0.05	4.00	False	False
’min cluster size’: 5, ’min samples’: 10	0.92	0.92	0.62	0.05	0.13	6.00	True	True
’min cluster size’: 5, ’min samples’: 20	0.93	0.93	0.61	0.08	0.14	6.00	True	True
’min cluster size’: 5, ’min samples’: 30	0.95	0.95	0.64	0.04	0.06	6.00	True	True
’min cluster size’: 5, ’min samples’: 40	0.72	0.72	0.55	0.07	0.04	4.00	False	False
’min cluster size’: 10, ’min samples’: 5	0.89	0.89	0.49	0.07	0.12	7.00	False	False
’min cluster size’: 10, ’min samples’: 7	0.91	0.91	0.61	0.06	0.21	6.00	True	True
’min cluster size’: 10, ’min samples’: 10	0.92	0.92	0.62	0.05	0.13	6.00	False	False
’min cluster size’: 10, ’min samples’: 20	0.93	0.93	0.61	0.08	0.14	6.00	False	False
’min cluster size’: 10, ’min samples’: 30	0.95	0.95	0.64	0.04	0.06	6.00	False	False
’min cluster size’: 10, ’min samples’: 40	0.72	0.72	0.55	0.07	0.04	4.00	False	False
’min cluster size’: 15, ’min samples’: 5	0.90	0.90	0.60	0.06	0.23	6.00	True	True
’min cluster size’: 15, ’min samples’: 7	0.91	0.91	0.61	0.06	0.21	6.00	False	False
’min cluster size’: 15, ’min samples’: 10	0.92	0.92	0.62	0.05	0.13	6.00	False	False
’min cluster size’: 15, ’min samples’: 20	0.93	0.93	0.61	0.08	0.14	6.00	False	False
’min cluster size’: 15, ’min samples’: 30	0.95	0.95	0.64	0.04	0.06	6.00	False	False
’min cluster size’: 15, ’min samples’: 40	0.72	0.72	0.55	0.07	0.04	4.00	False	False
’min cluster size’: 20, ’min samples’: 5	0.90	0.90	0.60	0.06	0.23	6.00	False	False
’min cluster size’: 20, ’min samples’: 7	0.91	0.91	0.61	0.06	0.21	6.00	False	False
’min cluster size’: 20, ’min samples’: 10	0.92	0.92	0.62	0.05	0.13	6.00	False	False
’min cluster size’: 20, ’min samples’: 20	0.93	0.93	0.61	0.08	0.14	6.00	False	False
’min cluster size’: 20, ’min samples’: 30	0.95	0.95	0.64	0.04	0.06	6.00	False	False
’min cluster size’: 20, ’min samples’: 40	0.72	0.72	0.55	0.07	0.04	4.00	False	False

Table C.83: Synthetic dataset - HDBSCAN clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
min cluster size’: 5, ’min samples’: 5	0.97	0.97	0.75	0.28	0.78	5.00	True	True
’min cluster size’: 5, ’min samples’: 7	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 5, ’min samples’: 10	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 5, ’min samples’: 20	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 5, ’min samples’: 30	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 5, ’min samples’: 40	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 10, ’min samples’: 5	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 10, ’min samples’: 7	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 10, ’min samples’: 10	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 10, ’min samples’: 20	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 10, ’min samples’: 30	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 10, ’min samples’: 40	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 15, ’min samples’: 5	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 15, ’min samples’: 7	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 15, ’min samples’: 10	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 15, ’min samples’: 20	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 15, ’min samples’: 30	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 15, ’min samples’: 40	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 20, ’min samples’: 5	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 20, ’min samples’: 7	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 20, ’min samples’: 10	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 20, ’min samples’: 20	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 20, ’min samples’: 30	0.97	0.97	0.75	0.28	0.78	5.00	False	False
’min cluster size’: 20, ’min samples’: 40	0.97	0.97	0.75	0.28	0.78	5.00	False	False

Table C.84: Synthetic dataset - HDBSCAN clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’min cluster size’: 5, ’min samples’: 5	0.41	0.43	0.31	0.09	0.44	146.00	False	False
’min cluster size’: 5, ’min samples’: 7	0.49	0.50	0.40	0.05	0.55	101.00	False	False
’min cluster size’: 5, ’min samples’: 10	0.70	0.71	0.62	0.10	0.72	41.00	False	False
’min cluster size’: 5, ’min samples’: 20	0.97	0.97	0.85	0.43	0.91	5.00	True	True
’min cluster size’: 5, ’min samples’: 30	0.97	0.97	0.85	0.43	0.91	5.00	False	False
’min cluster size’: 5, ’min samples’: 40	0.97	0.97	0.85	0.43	0.91	5.00	False	False
’min cluster size’: 10, ’min samples’: 5	0.51	0.53	0.45	0.07	0.51	86.00	False	False
’min cluster size’: 10, ’min samples’: 7	0.51	0.53	0.44	0.05	0.53	80.00	False	False
’min cluster size’: 10, ’min samples’: 10	0.83	0.83	0.74	0.08	0.81	23.00	False	False
’min cluster size’: 10, ’min samples’: 20	0.97	0.97	0.85	0.43	0.91	5.00	False	False
’min cluster size’: 10, ’min samples’: 30	0.97	0.97	0.85	0.43	0.91	5.00	False	False
’min cluster size’: 10, ’min samples’: 40	0.97	0.97	0.85	0.43	0.91	5.00	False	False
’min cluster size’: 15, ’min samples’: 5	0.55	0.55	0.48	0.08	0.47	64.00	False	False
’min cluster size’: 15, ’min samples’: 7	0.72	0.73	0.64	0.09	0.70	33.00	False	False
’min cluster size’: 15, ’min samples’: 10	0.84	0.84	0.76	0.13	0.81	16.00	False	False
’min cluster size’: 15, ’min samples’: 20	0.97	0.97	0.85	0.43	0.91	5.00	False	False
’min cluster size’: 15, ’min samples’: 30	0.97	0.97	0.85	0.43	0.91	5.00	False	False
’min cluster size’: 15, ’min samples’: 40	0.97	0.97	0.85	0.43	0.91	5.00	False	False
’min cluster size’: 20, ’min samples’: 5	0.84	0.84	0.76	0.12	0.79	16.00	False	False
’min cluster size’: 20, ’min samples’: 7	0.84	0.84	0.76	0.12	0.81	16.00	False	False
’min cluster size’: 20, ’min samples’: 10	0.84	0.84	0.76	0.13	0.81	16.00	False	False
’min cluster size’: 20, ’min samples’: 20	0.97	0.97	0.85	0.43	0.91	5.00	False	False
’min cluster size’: 20, ’min samples’: 30	0.97	0.97	0.85	0.43	0.91	5.00	False	False
’min cluster size’: 20, ’min samples’: 40	0.97	0.97	0.85	0.43	0.91	5.00	False	False

Table C.85: TEP dataset - Hierarchical clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’linkage’: ’ward’, ’n clusters’: 2	0.16	0.16	0.51	0.08	-0.54	2.00	False	False
’linkage’: ’ward’, ’n clusters’: 3	0.44	0.44	0.24	0.06	-0.60	3.00	False	False
’linkage’: ’ward’, ’n clusters’: 4	0.43	0.43	0.25	0.07	-0.58	4.00	False	False
’linkage’: ’ward’, ’n clusters’: 5	0.45	0.45	0.26	0.08	-0.51	5.00	False	False
’linkage’: ’ward’, ’n clusters’: 6	0.55	0.55	0.29	0.08	-0.34	6.00	False	False
’linkage’: ’ward’, ’n clusters’: 7	0.61	0.61	0.29	0.09	-0.26	7.00	False	False
’linkage’: ’ward’, ’n clusters’: 8	0.60	0.60	0.30	0.09	-0.25	8.00	False	False
’linkage’: ’ward’, ’n clusters’: 9	0.59	0.59	0.31	0.09	-0.25	9.00	False	False
’linkage’: ’ward’, ’n clusters’: 10	0.60	0.60	0.23	0.08	-0.26	10.00	False	False
’linkage’: ’single’, ’n clusters’: 2	0.44	0.44	0.37	0.03	0.11	2.00	False	False
’linkage’: ’single’, ’n clusters’: 3	0.42	0.42	0.39	0.04	0.14	3.00	True	True
’linkage’: ’single’, ’n clusters’: 4	0.42	0.42	0.36	0.05	0.00	4.00	True	False
’linkage’: ’single’, ’n clusters’: 5	0.41	0.41	0.35	0.06	0.00	5.00	True	False
’linkage’: ’single’, ’n clusters’: 6	0.41	0.41	0.10	0.06	0.00	6.00	True	False
’linkage’: ’single’, ’n clusters’: 7	0.41	0.41	0.09	0.06	0.00	7.00	False	False
’linkage’: ’single’, ’n clusters’: 8	0.46	0.46	0.10	0.07	0.00	8.00	True	False
’linkage’: ’single’, ’n clusters’: 9	0.46	0.46	0.05	0.07	0.00	9.00	True	False
’linkage’: ’single’, ’n clusters’: 10	0.46	0.46	-0.00	0.07	0.00	10.00	True	False
’linkage’: ’complete’, ’n clusters’: 2	0.21	0.21	0.45	0.07	-0.53	2.00	False	False
’linkage’: ’complete’, ’n clusters’: 3	0.29	0.29	0.44	0.07	-0.52	3.00	False	False
’linkage’: ’complete’, ’n clusters’: 4	0.28	0.28	0.40	0.10	-0.51	4.00	False	False
’linkage’: ’complete’, ’n clusters’: 5	0.27	0.27	0.41	0.10	-0.51	5.00	False	False
’linkage’: ’complete’, ’n clusters’: 6	0.27	0.27	0.39	0.11	-0.50	6.00	False	False
’linkage’: ’complete’, ’n clusters’: 7	0.33	0.33	0.41	0.11	-0.49	7.00	False	False
’linkage’: ’complete’, ’n clusters’: 8	0.37	0.37	0.33	0.11	-0.20	8.00	True	True
’linkage’: ’complete’, ’n clusters’: 9	0.36	0.37	0.33	0.10	-0.20	9.00	False	False
’linkage’: ’complete’, ’n clusters’: 10	0.36	0.36	0.33	0.10	-0.20	10.00	False	False
’linkage’: ’average’, ’n clusters’: 2	0.07	0.07	0.52	0.13	-0.49	2.00	True	True
’linkage’: ’average’, ’n clusters’: 3	0.16	0.16	0.49	0.10	-0.47	3.00	True	True
’linkage’: ’average’, ’n clusters’: 4	0.16	0.16	0.48	0.12	-0.46	4.00	True	True
’linkage’: ’average’, ’n clusters’: 5	0.23	0.23	0.40	0.11	-0.51	5.00	False	False
’linkage’: ’average’, ’n clusters’: 6	0.30	0.30	0.39	0.10	-0.49	6.00	False	False
’linkage’: ’average’, ’n clusters’: 7	0.41	0.42	0.40	0.10	-0.02	7.00	True	True
’linkage’: ’average’, ’n clusters’: 8	0.41	0.41	0.40	0.10	-0.01	8.00	True	True
’linkage’: ’average’, ’n clusters’: 9	0.41	0.41	0.35	0.10	-0.01	9.00	True	True
’linkage’: ’average’, ’n clusters’: 10	0.40	0.40	0.32	0.10	-0.01	10.00	True	False

Table C.86: TEP dataset - Hierarchical clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'linkage': 'ward', 'n clusters': 2	0.24	0.24	0.69	0.09	-0.73	2.00	False	False
'linkage': 'ward', 'n clusters': 3	0.23	0.23	0.70	0.14	-0.72	3.00	False	False
'linkage': 'ward', 'n clusters': 4	0.42	0.42	0.58	0.14	-0.77	4.00	False	False
'linkage': 'ward', 'n clusters': 5	0.51	0.51	0.68	0.16	-0.47	5.00	False	False
'linkage': 'ward', 'n clusters': 6	0.50	0.50	0.69	0.19	-0.46	6.00	True	True
'linkage': 'ward', 'n clusters': 7	0.49	0.50	0.69	0.18	-0.46	7.00	False	False
'linkage': 'ward', 'n clusters': 8	0.51	0.51	0.70	0.18	-0.42	8.00	False	False
'linkage': 'ward', 'n clusters': 9	0.50	0.50	0.70	0.18	-0.41	9.00	True	True
'linkage': 'ward', 'n clusters': 10	0.50	0.50	0.70	0.18	-0.41	10.00	True	True
'linkage': 'single', 'n clusters': 2	0.37	0.37	0.61	0.04	-0.39	2.00	False	False
'linkage': 'single', 'n clusters': 3	0.37	0.37	0.45	0.05	0.00	3.00	True	False
'linkage': 'single', 'n clusters': 4	0.37	0.37	0.42	0.06	0.00	4.00	False	False
'linkage': 'single', 'n clusters': 5	0.37	0.37	0.39	0.07	0.00	5.00	False	False
'linkage': 'single', 'n clusters': 6	0.37	0.37	0.38	0.07	0.00	6.00	False	False
'linkage': 'single', 'n clusters': 7	0.37	0.37	0.38	0.08	0.00	7.00	False	False
'linkage': 'single', 'n clusters': 8	0.37	0.37	0.38	0.08	0.00	8.00	False	False
'linkage': 'single', 'n clusters': 9	0.34	0.34	0.44	0.09	0.00	9.00	True	False
'linkage': 'single', 'n clusters': 10	0.34	0.34	0.43	0.09	0.00	10.00	True	False
'linkage': 'complete', 'n clusters': 2	0.23	0.23	0.49	0.04	-0.96	2.00	False	False
'linkage': 'complete', 'n clusters': 3	0.27	0.27	0.50	0.09	-0.93	3.00	False	False
'linkage': 'complete', 'n clusters': 4	0.27	0.27	0.51	0.11	-0.91	4.00	False	False
'linkage': 'complete', 'n clusters': 5	0.29	0.29	0.46	0.14	-0.90	5.00	False	False
'linkage': 'complete', 'n clusters': 6	0.29	0.29	0.46	0.15	-0.90	6.00	False	False
'linkage': 'complete', 'n clusters': 7	0.33	0.33	0.50	0.17	-0.87	7.00	False	False
'linkage': 'complete', 'n clusters': 8	0.32	0.32	0.50	0.17	-0.88	8.00	False	False
'linkage': 'complete', 'n clusters': 9	0.38	0.38	0.55	0.17	-0.87	9.00	False	False
'linkage': 'complete', 'n clusters': 10	0.37	0.38	0.55	0.16	-0.86	10.00	False	False
'linkage': 'average', 'n clusters': 2	0.15	0.15	0.71	0.15	-0.48	2.00	True	True
'linkage': 'average', 'n clusters': 3	0.14	0.14	0.70	0.18	-0.54	3.00	False	False
'linkage': 'average', 'n clusters': 4	0.21	0.21	0.68	0.18	-0.63	4.00	False	False
'linkage': 'average', 'n clusters': 5	0.33	0.33	0.66	0.16	-0.39	5.00	True	True
'linkage': 'average', 'n clusters': 6	0.36	0.36	0.62	0.17	-0.32	6.00	True	True
'linkage': 'average', 'n clusters': 7	0.36	0.36	0.62	0.17	-0.33	7.00	True	True
'linkage': 'average', 'n clusters': 8	0.35	0.35	0.61	0.17	-0.33	8.00	False	False
'linkage': 'average', 'n clusters': 9	0.35	0.35	0.60	0.17	-0.32	9.00	True	True
'linkage': 'average', 'n clusters': 10	0.34	0.34	0.60	0.18	-0.32	10.00	True	True

Table C.87: TEP dataset - Hierarchical clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’linkage’: ’ward’, ’n clusters’: 7	0.51	0.51	0.40	0.13	-0.68	7.00	False	False
’linkage’: ’ward’, ’n clusters’: 8	0.50	0.51	0.40	0.14	-0.67	8.00	False	False
’linkage’: ’ward’, ’n clusters’: 9	0.50	0.50	0.41	0.15	-0.67	9.00	False	False
’linkage’: ’ward’, ’n clusters’: 10	0.47	0.47	0.39	0.14	-0.72	10.00	False	False
’linkage’: ’single’, ’n clusters’: 2	-0.00	0.00	0.52	0.32	0.00	2.00	True	False
’linkage’: ’single’, ’n clusters’: 3	0.00	0.00	0.45	0.31	0.00	3.00	False	False
’linkage’: ’single’, ’n clusters’: 4	0.00	0.00	0.45	0.31	0.00	4.00	False	False
’linkage’: ’single’, ’n clusters’: 5	0.00	0.00	0.06	0.23	0.00	5.00	False	False
’linkage’: ’single’, ’n clusters’: 6	0.00	0.00	0.06	0.24	0.00	6.00	False	False
’linkage’: ’single’, ’n clusters’: 7	0.00	0.00	0.05	0.24	0.00	7.00	False	False
’linkage’: ’single’, ’n clusters’: 8	0.00	0.00	-0.12	0.22	0.00	8.00	False	False
’linkage’: ’single’, ’n clusters’: 9	0.44	0.44	0.02	0.07	0.00	9.00	False	False
’linkage’: ’single’, ’n clusters’: 10	0.45	0.45	0.03	0.07	0.00	10.00	False	False
’linkage’: ’complete’, ’n clusters’: 2	0.17	0.17	0.52	0.10	-0.76	2.00	False	False
’linkage’: ’complete’, ’n clusters’: 3	0.24	0.25	0.32	0.10	-0.87	3.00	False	False
’linkage’: ’complete’, ’n clusters’: 4	0.24	0.24	0.33	0.12	-0.86	4.00	False	False
’linkage’: ’complete’, ’n clusters’: 5	0.30	0.30	0.37	0.12	-0.84	5.00	False	False
’linkage’: ’complete’, ’n clusters’: 6	0.50	0.50	0.45	0.13	-0.67	6.00	False	False
’linkage’: ’complete’, ’n clusters’: 7	0.50	0.50	0.43	0.13	-0.66	7.00	False	False
’linkage’: ’complete’, ’n clusters’: 8	0.52	0.52	0.44	0.14	-0.66	8.00	False	False
’linkage’: ’complete’, ’n clusters’: 9	0.51	0.51	0.44	0.15	-0.65	9.00	False	False
’linkage’: ’complete’, ’n clusters’: 10	0.51	0.51	0.44	0.15	-0.65	10.00	False	False
’linkage’: ’average’, ’n clusters’: 2	0.10	0.10	0.52	0.15	-0.53	2.00	False	False
’linkage’: ’average’, ’n clusters’: 3	0.21	0.21	0.54	0.14	-0.70	3.00	True	True
’linkage’: ’average’, ’n clusters’: 4	0.25	0.25	0.46	0.15	-0.66	4.00	False	False
’linkage’: ’average’, ’n clusters’: 5	0.25	0.25	0.41	0.18	-0.67	5.00	False	False
’linkage’: ’average’, ’n clusters’: 6	0.38	0.38	0.42	0.15	-0.40	6.00	False	False
’linkage’: ’average’, ’n clusters’: 7	0.37	0.37	0.43	0.16	-0.39	7.00	False	False
’linkage’: ’average’, ’n clusters’: 8	0.36	0.36	0.38	0.18	-0.40	8.00	False	False
’linkage’: ’average’, ’n clusters’: 9	0.55	0.55	0.47	0.17	-0.39	9.00	False	False
’linkage’: ’average’, ’n clusters’: 10	0.54	0.54	0.46	0.16	-0.39	10.00	False	False

Table C.88: TEP dataset - Hierarchical clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'linkage': 'ward', 'n clusters': 2	0.37	0.37	0.67	0.09	-0.57	2.00	False	False
'linkage': 'ward', 'n clusters': 3	0.35	0.35	0.70	0.15	-0.51	3.00	True	True
'linkage': 'ward', 'n clusters': 4	0.33	0.34	0.68	0.17	-0.55	4.00	False	False
'linkage': 'ward', 'n clusters': 5	0.61	0.61	0.62	0.15	-0.30	5.00	False	False
'linkage': 'ward', 'n clusters': 6	0.66	0.66	0.66	0.18	-0.25	6.00	False	False
'linkage': 'ward', 'n clusters': 7	0.65	0.65	0.66	0.19	-0.25	7.00	False	False
'linkage': 'ward', 'n clusters': 8	0.64	0.64	0.67	0.20	-0.22	8.00	False	False
'linkage': 'ward', 'n clusters': 9	0.61	0.61	0.65	0.20	-0.35	9.00	False	False
'linkage': 'ward', 'n clusters': 10	0.60	0.60	0.65	0.18	-0.37	10.00	False	False
'linkage': 'single', 'n clusters': 2	0.25	0.25	0.61	0.11	-0.63	2.00	False	False
'linkage': 'single', 'n clusters': 3	0.25	0.25	0.56	0.14	0.00	3.00	False	False
'linkage': 'single', 'n clusters': 4	0.24	0.24	0.56	0.17	0.00	4.00	False	False
'linkage': 'single', 'n clusters': 5	0.24	0.24	0.20	0.15	0.00	5.00	False	False
'linkage': 'single', 'n clusters': 6	0.34	0.34	0.41	0.15	0.00	6.00	False	False
'linkage': 'single', 'n clusters': 7	0.34	0.34	0.39	0.16	0.00	7.00	False	False
'linkage': 'single', 'n clusters': 8	0.34	0.34	0.38	0.16	0.00	8.00	False	False
'linkage': 'single', 'n clusters': 9	0.34	0.34	0.38	0.17	0.00	9.00	False	False
'linkage': 'single', 'n clusters': 10	0.34	0.34	0.37	0.15	0.00	10.00	False	False
'linkage': 'complete', 'n clusters': 2	0.27	0.27	0.70	0.11	-0.71	2.00	False	False
'linkage': 'complete', 'n clusters': 3	0.26	0.26	0.67	0.13	-0.71	3.00	False	False
'linkage': 'complete', 'n clusters': 4	0.25	0.25	0.67	0.18	-0.68	4.00	False	False
'linkage': 'complete', 'n clusters': 5	0.25	0.25	0.66	0.18	-0.67	5.00	False	False
'linkage': 'complete', 'n clusters': 6	0.40	0.40	0.69	0.18	-0.49	6.00	True	True
'linkage': 'complete', 'n clusters': 7	0.40	0.40	0.68	0.19	-0.47	7.00	False	False
'linkage': 'complete', 'n clusters': 8	0.39	0.39	0.68	0.19	-0.47	8.00	True	True
'linkage': 'complete', 'n clusters': 9	0.39	0.39	0.68	0.20	-0.48	9.00	True	True
'linkage': 'complete', 'n clusters': 10	0.38	0.38	0.67	0.19	-0.48	10.00	False	False
'linkage': 'average', 'n clusters': 2	0.14	0.14	0.74	0.18	-0.56	2.00	True	True
'linkage': 'average', 'n clusters': 3	0.14	0.14	0.72	0.19	-0.67	3.00	True	True
'linkage': 'average', 'n clusters': 4	0.34	0.34	0.69	0.17	-0.55	4.00	False	False
'linkage': 'average', 'n clusters': 5	0.33	0.33	0.69	0.19	-0.55	5.00	True	True
'linkage': 'average', 'n clusters': 6	0.42	0.42	0.68	0.19	0.21	6.00	True	True
'linkage': 'average', 'n clusters': 7	0.41	0.41	0.68	0.20	0.23	7.00	True	True
'linkage': 'average', 'n clusters': 8	0.40	0.41	0.68	0.21	0.24	8.00	True	True
'linkage': 'average', 'n clusters': 9	0.40	0.40	0.68	0.20	0.26	9.00	True	True
'linkage': 'average', 'n clusters': 10	0.40	0.40	0.55	0.20	0.24	10.00	False	False

Table C.89: TEP dataset - Hierarchical clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’linkage’: ’ward’, ’n clusters’: 2	0.52	0.53	0.43	0.13	-0.64	2.00	False	False
’linkage’: ’ward’, ’n clusters’: 3	0.67	0.67	0.46	0.12	-0.63	3.00	False	False
’linkage’: ’ward’, ’n clusters’: 4	0.65	0.65	0.48	0.14	-0.63	4.00	False	False
’linkage’: ’ward’, ’n clusters’: 5	0.72	0.72	0.45	0.14	-0.58	5.00	False	False
’linkage’: ’ward’, ’n clusters’: 6	0.71	0.71	0.47	0.14	-0.53	6.00	False	False
’linkage’: ’ward’, ’n clusters’: 7	0.76	0.76	0.50	0.16	-0.47	7.00	False	False
’linkage’: ’ward’, ’n clusters’: 8	0.73	0.73	0.50	0.16	-0.45	8.00	True	True
’linkage’: ’ward’, ’n clusters’: 9	0.71	0.71	0.50	0.17	-0.45	9.00	True	True
’linkage’: ’ward’, ’n clusters’: 10	0.68	0.68	0.48	0.14	-0.48	10.00	False	False
’linkage’: ’single’, ’n clusters’: 2	0.53	0.53	0.38	0.09	0.10	2.00	True	True
’linkage’: ’single’, ’n clusters’: 3	0.77	0.77	0.45	0.10	-0.23	3.00	True	True
’linkage’: ’single’, ’n clusters’: 4	0.75	0.75	0.23	0.11	-0.24	4.00	True	True
’linkage’: ’single’, ’n clusters’: 5	0.76	0.76	0.17	0.12	-0.28	5.00	True	True
’linkage’: ’single’, ’n clusters’: 6	0.76	0.76	0.07	0.10	-0.30	6.00	False	False
’linkage’: ’single’, ’n clusters’: 7	0.76	0.76	-0.00	0.10	-0.31	7.00	False	False
’linkage’: ’single’, ’n clusters’: 8	0.76	0.76	-0.05	0.10	-0.30	8.00	False	False
’linkage’: ’single’, ’n clusters’: 9	0.71	0.71	0.05	0.13	-0.27	9.00	True	True
’linkage’: ’single’, ’n clusters’: 10	0.71	0.71	-0.05	0.13	0.00	10.00	True	False
’linkage’: ’complete’, ’n clusters’: 2	0.52	0.53	0.43	0.13	-0.64	2.00	False	False
’linkage’: ’complete’, ’n clusters’: 3	0.60	0.60	0.42	0.13	-0.69	3.00	False	False
’linkage’: ’complete’, ’n clusters’: 4	0.73	0.73	0.49	0.15	-0.60	4.00	False	False
’linkage’: ’complete’, ’n clusters’: 5	0.65	0.65	0.46	0.13	-0.70	5.00	False	False
’linkage’: ’complete’, ’n clusters’: 6	0.63	0.63	0.47	0.15	-0.63	6.00	False	False
’linkage’: ’complete’, ’n clusters’: 7	0.60	0.60	0.44	0.15	-0.67	7.00	False	False
’linkage’: ’complete’, ’n clusters’: 8	0.60	0.60	0.40	0.13	-0.67	8.00	False	False
’linkage’: ’complete’, ’n clusters’: 9	0.59	0.59	0.40	0.13	-0.64	9.00	False	False
’linkage’: ’complete’, ’n clusters’: 10	0.58	0.58	0.38	0.15	-0.63	10.00	False	False
’linkage’: ’average’, ’n clusters’: 2	0.52	0.53	0.43	0.13	-0.64	2.00	False	False
’linkage’: ’average’, ’n clusters’: 3	0.55	0.55	0.37	0.14	-0.67	3.00	False	False
’linkage’: ’average’, ’n clusters’: 4	0.74	0.74	0.47	0.13	-0.49	4.00	False	False
’linkage’: ’average’, ’n clusters’: 5	0.71	0.71	0.47	0.16	-0.49	5.00	False	False
’linkage’: ’average’, ’n clusters’: 6	0.71	0.71	0.47	0.16	-0.49	6.00	False	False
’linkage’: ’average’, ’n clusters’: 7	0.67	0.67	0.46	0.17	-0.47	7.00	False	False
’linkage’: ’average’, ’n clusters’: 8	0.72	0.72	0.50	0.16	-0.52	8.00	False	False
’linkage’: ’average’, ’n clusters’: 9	0.71	0.71	0.50	0.17	-0.52	9.00	False	False
’linkage’: ’average’, ’n clusters’: 10	0.70	0.70	0.50	0.17	-0.50	10.00	True	True

Table C.90: TEP dataset - Hierarchical clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’linkage’: ’ward’, ’n clusters’: 2	0.45	0.45	0.43	0.14	-0.66	2.00	False	False
’linkage’: ’ward’, ’n clusters’: 3	0.59	0.59	0.57	0.15	-0.60	3.00	False	False
’linkage’: ’ward’, ’n clusters’: 4	0.62	0.62	0.50	0.13	-0.58	4.00	False	False
’linkage’: ’ward’, ’n clusters’: 5	0.60	0.60	0.52	0.15	-0.56	5.00	False	False
’linkage’: ’ward’, ’n clusters’: 6	0.64	0.64	0.47	0.15	-0.54	6.00	False	False
’linkage’: ’ward’, ’n clusters’: 7	0.64	0.64	0.49	0.15	-0.52	7.00	False	False
’linkage’: ’ward’, ’n clusters’: 8	0.62	0.62	0.46	0.14	-0.49	8.00	False	False
’linkage’: ’ward’, ’n clusters’: 9	0.60	0.61	0.44	0.14	-0.50	9.00	False	False
’linkage’: ’ward’, ’n clusters’: 10	0.60	0.60	0.44	0.14	-0.52	10.00	False	False
’linkage’: ’single’, ’n clusters’: 2	0.52	0.52	0.46	0.08	-0.29	2.00	True	True
’linkage’: ’single’, ’n clusters’: 3	0.48	0.48	0.42	0.13	-0.33	3.00	True	True
’linkage’: ’single’, ’n clusters’: 4	0.48	0.48	0.17	0.13	-0.39	4.00	True	True
’linkage’: ’single’, ’n clusters’: 5	0.48	0.48	0.10	0.13	0.00	5.00	True	False
’linkage’: ’single’, ’n clusters’: 6	0.48	0.48	0.04	0.13	0.00	6.00	False	False
’linkage’: ’single’, ’n clusters’: 7	0.48	0.48	0.04	0.13	0.00	7.00	False	False
’linkage’: ’single’, ’n clusters’: 8	0.46	0.46	0.02	0.13	0.00	8.00	False	False
’linkage’: ’single’, ’n clusters’: 9	0.46	0.46	0.01	0.13	0.00	9.00	False	False
’linkage’: ’single’, ’n clusters’: 10	0.46	0.46	-0.16	0.14	0.00	10.00	True	False
’linkage’: ’complete’, ’n clusters’: 2	0.39	0.39	0.35	0.07	-0.87	2.00	False	False
’linkage’: ’complete’, ’n clusters’: 3	0.44	0.44	0.36	0.12	-0.79	3.00	False	False
’linkage’: ’complete’, ’n clusters’: 4	0.42	0.42	0.35	0.13	-0.79	4.00	False	False
’linkage’: ’complete’, ’n clusters’: 5	0.55	0.55	0.50	0.14	-0.72	5.00	False	False
’linkage’: ’complete’, ’n clusters’: 6	0.57	0.57	0.46	0.15	-0.68	6.00	False	False
’linkage’: ’complete’, ’n clusters’: 7	0.56	0.56	0.46	0.15	-0.68	7.00	False	False
’linkage’: ’complete’, ’n clusters’: 8	0.61	0.61	0.41	0.14	-0.66	8.00	False	False
’linkage’: ’complete’, ’n clusters’: 9	0.59	0.59	0.43	0.14	-0.61	9.00	False	False
’linkage’: ’complete’, ’n clusters’: 10	0.59	0.59	0.43	0.14	-0.63	10.00	False	False
’linkage’: ’average’, ’n clusters’: 2	0.44	0.44	0.49	0.09	-0.47	2.00	True	True
’linkage’: ’average’, ’n clusters’: 3	0.41	0.41	0.47	0.14	-0.51	3.00	False	False
’linkage’: ’average’, ’n clusters’: 4	0.59	0.59	0.59	0.17	-0.51	4.00	True	True
’linkage’: ’average’, ’n clusters’: 5	0.57	0.57	0.58	0.19	-0.49	5.00	True	True
’linkage’: ’average’, ’n clusters’: 6	0.56	0.56	0.57	0.19	-0.52	6.00	True	True
’linkage’: ’average’, ’n clusters’: 7	0.58	0.58	0.48	0.17	-0.56	7.00	False	False
’linkage’: ’average’, ’n clusters’: 8	0.62	0.62	0.41	0.16	-0.59	8.00	False	False
’linkage’: ’average’, ’n clusters’: 9	0.61	0.61	0.41	0.15	-0.58	9.00	False	False
’linkage’: ’average’, ’n clusters’: 10	0.63	0.63	0.44	0.16	-0.51	10.00	False	False

Table C.91: TEP dataset - k-means clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.16	0.16	0.51	0.08	-0.51	2.00	True	True
’n clusters’: 3	0.42	0.42	0.25	0.06	-0.64	3.00	False	False
’n clusters’: 4	0.43	0.43	0.26	0.07	-0.63	4.00	False	False
’n clusters’: 5	0.49	0.49	0.28	0.08	-0.57	5.00	False	False
’n clusters’: 6	0.54	0.54	0.30	0.08	-0.58	6.00	False	False
’n clusters’: 7	0.58	0.58	0.31	0.09	-0.51	7.00	True	True
’n clusters’: 8	0.57	0.57	0.31	0.10	-0.52	8.00	True	True
’n clusters’: 9	0.59	0.59	0.23	0.08	-0.49	9.00	False	False
’n clusters’: 10	0.59	0.58	0.24	0.08	-0.37	10.00	True	True
’n clusters’: 11	0.59	0.61	0.22	0.08	-0.42	11.00	False	False

Table C.92: TEP dataset - k-means clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.17	0.18	0.72	0.13	-0.47	2.00	True	True
’n clusters’: 3	0.21	0.21	0.56	0.16	-0.65	3.00	False	False
’n clusters’: 4	0.41	0.41	0.60	0.14	-0.84	4.00	False	False
’n clusters’: 5	0.51	0.51	0.68	0.16	-0.78	5.00	False	False
’n clusters’: 6	0.50	0.50	0.70	0.19	-0.42	6.00	True	True
’n clusters’: 7	0.50	0.50	0.70	0.18	-0.55	7.00	False	False
’n clusters’: 8	0.50	0.50	0.71	0.19	-0.52	8.00	True	True
’n clusters’: 9	0.50	0.50	0.71	0.19	-0.54	9.00	True	True
’n clusters’: 10	0.49	0.50	0.71	0.19	-0.54	10.00	True	True
’n clusters’: 11	0.49	0.49	0.71	0.19	-0.54	11.00	True	True

Table C.93: TEP dataset - *k*-means clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
n clusters’: 2	0.18	0.18	0.57	0.11	-0.74	2.00	True	True
’n clusters’: 3	0.44	0.44	0.42	0.11	-0.70	3.00	False	False
’n clusters’: 4	0.47	0.47	0.45	0.12	-0.72	4.00	False	False
’n clusters’: 5	0.51	0.51	0.45	0.13	-0.63	5.00	True	True
’n clusters’: 6	0.46	0.47	0.41	0.13	-0.77	6.00	False	False
’n clusters’: 7	0.50	0.50	0.43	0.14	-0.67	7.00	True	True
’n clusters’: 8	0.51	0.51	0.43	0.15	-0.69	8.00	True	True
’n clusters’: 9	0.49	0.49	0.41	0.14	-0.73	9.00	False	False
’n clusters’: 10	0.49	0.49	0.41	0.14	-0.74	10.00	False	False
’n clusters’: 11	0.49	0.49	0.42	0.14	-0.74	11.00	False	False

Table C.94: TEP dataset - *k*-means clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.19	0.19	0.74	0.14	-0.75	2.00	True	True
’n clusters’: 3	0.35	0.35	0.70	0.16	-0.77	3.00	False	False
’n clusters’: 4	0.40	0.40	0.71	0.16	-0.54	4.00	True	True
’n clusters’: 5	0.59	0.57	0.66	0.16	-0.63	5.00	False	False
’n clusters’: 6	0.59	0.59	0.68	0.20	-0.54	6.00	True	True
’n clusters’: 7	0.58	0.58	0.69	0.20	-0.55	7.00	True	True
’n clusters’: 8	0.57	0.57	0.70	0.20	-0.56	8.00	True	True
’n clusters’: 9	0.59	0.59	0.67	0.19	-0.28	9.00	True	True
’n clusters’: 10	0.59	0.58	0.67	0.19	-0.31	10.00	True	True
’n clusters’: 11	0.58	0.58	0.66	0.18	-0.26	11.00	True	True

Table C.95: TEP dataset - *k*-means clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.19	0.19	0.74	0.14	-0.75	2.00	True	True
’n clusters’: 3	0.35	0.35	0.70	0.16	-0.77	3.00	False	False
’n clusters’: 4	0.40	0.40	0.71	0.16	-0.54	4.00	True	True
’n clusters’: 5	0.59	0.59	0.66	0.16	-0.63	5.00	False	False
’n clusters’: 6	0.59	0.59	0.68	0.20	-0.55	6.00	True	True
’n clusters’: 7	0.58	0.58	0.69	0.20	-0.54	7.00	True	True
’n clusters’: 8	0.57	0.57	0.70	0.20	-0.56	8.00	True	True
’n clusters’: 9	0.59	0.59	0.67	0.19	-0.28	9.00	True	True
’n clusters’: 10	0.58	0.59	0.67	0.19	-0.31	10.00	True	True
’n clusters’: 11	0.57	0.58	0.66	0.18	-0.28	11.00	False	False

Table C.96: TEP dataset - *k*-means clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.48	0.48	0.43	0.13	-0.75	2.00	False	False
’n clusters’: 3	0.61	0.61	0.47	0.12	-0.86	3.00	False	False
’n clusters’: 4	0.65	0.65	0.51	0.14	-0.89	4.00	False	False
’n clusters’: 5	0.61	0.61	0.47	0.13	-0.80	5.00	False	False
’n clusters’: 6	0.75	0.75	0.48	0.14	-0.71	6.00	False	False
’n clusters’: 7	0.73	0.72	0.49	0.15	-0.69	7.00	False	False
’n clusters’: 8	0.70	0.71	0.51	0.16	-0.65	8.00	False	False
’n clusters’: 9	0.69	0.69	0.51	0.17	-0.61	9.00	True	True
’n clusters’: 10	0.66	0.66	0.51	0.15	-0.68	10.00	False	False
’n clusters’: 11	0.64	0.65	0.51	0.16	-0.56	11.00	True	True

Table C.97: TEP dataset - *k*-medoids clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'n clusters': 2	0.00	0.00	0.00	0.00	nan	1.00	True	False
'n clusters': 3	0.00	0.00	0.00	0.00	nan	1.00	False	False
'n clusters': 4	0.00	0.00	0.00	0.00	nan	1.00	False	False
'n clusters': 5	0.34	0.34	0.18	0.05	-0.70	5.00	True	True
'n clusters': 6	0.28	0.28	0.04	0.03	-0.79	6.00	False	False
'n clusters': 7	0.30	0.30	0.02	0.03	-0.76	7.00	False	False
'n clusters': 8	0.27	0.27	0.01	0.03	-0.78	8.00	False	False
'n clusters': 9	0.27	0.27	-0.01	0.03	-0.78	9.00	False	False
'n clusters': 10	0.33	0.33	0.00	0.03	-0.73	10.00	False	False
'n clusters': 11	0.33	0.33	-0.01	0.03	-0.72	11.00	False	False

Table C.98: TEP dataset - *k*-medoids clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'n clusters': 2	0.29	0.29	0.42	0.10	-0.87	2.00	True	True
'n clusters': 3	0.23	0.23	0.19	0.06	-0.96	3.00	False	False
'n clusters': 4	0.26	0.26	0.15	0.06	-0.97	4.00	False	False
'n clusters': 5	0.38	0.38	0.37	0.08	-0.90	5.00	False	False
'n clusters': 6	0.46	0.46	0.28	0.09	-0.86	6.00	False	False
'n clusters': 7	0.38	0.38	0.19	0.08	-0.88	7.00	False	False
'n clusters': 8	0.38	0.38	0.24	0.07	-0.88	8.00	False	False
'n clusters': 9	0.36	0.36	0.16	0.07	-0.85	9.00	False	False
'n clusters': 10	0.42	0.42	0.32	0.11	-0.77	10.00	True	True
'n clusters': 11	0.42	0.42	0.33	0.11	-0.74	11.00	True	True

Table C.99: TEP dataset - *k*-medoids clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'n clusters': 2	0.38	0.38	0.35	0.09	-0.82	2.00	False	False
'n clusters': 3	0.29	0.29	0.31	0.09	-0.91	3.00	False	False
'n clusters': 4	0.29	0.29	0.28	0.09	-0.92	4.00	False	False
'n clusters': 5	0.41	0.41	0.28	0.09	-0.74	5.00	False	False
'n clusters': 6	0.35	0.35	0.23	0.09	-0.83	6.00	False	False
'n clusters': 7	0.40	0.40	0.30	0.10	-0.76	7.00	False	False
'n clusters': 8	0.38	0.39	0.28	0.10	-0.77	8.00	False	False
'n clusters': 9	0.47	0.47	0.37	0.13	-0.68	9.00	True	True
'n clusters': 10	0.40	0.40	0.29	0.11	-0.78	10.00	False	False
'n clusters': 11	0.44	0.44	0.35	0.12	-0.73	11.00	False	False

Table C.100: TEP dataset - *k*-medoids clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'n clusters': 2	0.35	0.35	0.37	0.07	-0.89	2.00	False	False
'n clusters': 3	0.57	0.57	0.52	0.11	-0.84	3.00	False	False
'n clusters': 4	0.57	0.57	0.62	0.15	-0.63	4.00	True	True
'n clusters': 5	0.54	0.54	0.37	0.12	-0.83	5.00	False	False
'n clusters': 6	0.57	0.57	0.34	0.12	-0.81	6.00	False	False
'n clusters': 7	0.58	0.58	0.39	0.12	-0.82	7.00	False	False
'n clusters': 8	0.55	0.55	0.39	0.12	-0.80	8.00	False	False
'n clusters': 9	0.54	0.54	0.35	0.12	-0.79	9.00	False	False
'n clusters': 10	0.55	0.55	0.39	0.13	-0.71	10.00	False	False
'n clusters': 11	0.52	0.52	0.33	0.11	-0.76	11.00	False	False

Table C.101: TEP dataset - *k*-medoids clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.49	0.49	0.43	0.13	-0.90	2.00	False	False
’n clusters’: 3	0.62	0.62	0.46	0.12	-0.75	3.00	False	False
’n clusters’: 4	0.66	0.66	0.41	0.12	-0.85	4.00	False	False
’n clusters’: 5	0.68	0.68	0.45	0.13	-0.84	5.00	False	False
’n clusters’: 6	0.62	0.62	0.41	0.12	-0.81	6.00	False	False
’n clusters’: 7	0.72	0.72	0.48	0.15	-0.79	7.00	True	True
’n clusters’: 8	0.68	0.68	0.47	0.13	-0.72	8.00	False	False
’n clusters’: 9	0.65	0.65	0.47	0.14	-0.69	9.00	False	False
’n clusters’: 10	0.64	0.64	0.46	0.14	-0.70	10.00	False	False
’n clusters’: 11	0.64	0.64	0.49	0.15	-0.59	11.00	True	True

Table C.102: TEP dataset - *k*-medoids clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’n clusters’: 2	0.37	0.37	0.47	0.11	-0.76	2.00	False	False
’n clusters’: 3	0.60	0.60	0.57	0.15	-0.55	3.00	True	True
’n clusters’: 4	0.61	0.61	0.51	0.13	-0.59	4.00	False	False
’n clusters’: 5	0.61	0.61	0.43	0.13	-0.71	5.00	False	False
’n clusters’: 6	0.58	0.58	0.39	0.13	-0.66	6.00	False	False
’n clusters’: 7	0.58	0.58	0.38	0.13	-0.62	7.00	False	False
’n clusters’: 8	0.56	0.56	0.37	0.13	-0.64	8.00	False	False
’n clusters’: 9	0.55	0.55	0.35	0.12	-0.66	9.00	False	False
’n clusters’: 10	0.53	0.53	0.34	0.12	-0.64	10.00	False	False
’n clusters’: 11	0.53	0.53	0.33	0.12	-0.68	11.00	False	False

Table C.103: TEP dataset - GMM clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’covariance type’: ’full’, ’n ’: 11	0.63	0.68	0.14	0.02	-0.49	11.00	False	False
’covariance type’: ’tied’, ’n ’: 2	0.13	0.13	0.17	0.10	-0.57	2.00	True	True
’covariance type’: ’tied’, ’n ’: 3	0.46	0.46	0.22	0.08	-0.76	3.00	False	False
’covariance type’: ’tied’, ’n ’: 4	0.52	0.48	0.24	0.08	-0.76	4.00	False	False
’covariance type’: ’tied’, ’n ’: 5	0.59	0.57	0.25	0.07	-0.53	5.00	False	False
’covariance type’: ’tied’, ’n ’: 6	0.56	0.63	0.29	0.08	-0.66	6.00	False	False
’covariance type’: ’tied’, ’n ’: 7	0.64	0.63	0.30	0.09	-0.34	7.00	False	False
’covariance type’: ’tied’, ’n ’: 8	0.63	0.63	0.21	0.08	-0.41	8.00	False	False
’covariance type’: ’tied’, ’n ’: 9	0.65	0.61	0.23	0.08	-0.37	9.00	False	False
’covariance type’: ’tied’, ’n ’: 10	0.61	0.60	0.21	0.07	-0.36	10.00	False	False
’covariance type’: ’tied’, ’n ’: 11	0.62	0.60	0.22	0.07	-0.30	11.00	True	True
’covariance type’: ’diag’, ’n ’: 2	0.54	0.54	0.34	0.02	-0.36	2.00	False	False
’covariance type’: ’diag’, ’n ’: 3	0.49	0.63	0.33	0.05	-0.36	3.00	False	False
’covariance type’: ’diag’, ’n ’: 4	0.61	0.65	0.20	0.06	-0.46	4.00	False	False
’covariance type’: ’diag’, ’n ’: 5	0.63	0.63	0.23	0.07	-0.44	5.00	False	False
’covariance type’: ’diag’, ’n ’: 6	0.68	0.68	0.25	0.07	-0.38	6.00	False	False
’covariance type’: ’diag’, ’n ’: 7	0.66	0.67	0.28	0.08	-0.39	7.00	False	False
’covariance type’: ’diag’, ’n ’: 8	0.63	0.63	0.23	0.07	-0.45	8.00	False	False
’covariance type’: ’diag’, ’n ’: 9	0.62	0.62	0.20	0.04	-0.37	9.00	False	False
’covariance type’: ’diag’, ’n ’: 10	0.62	0.62	0.13	0.08	-0.36	10.00	False	False
’covariance type’: ’diag’, ’n ’: 11	0.61	0.56	0.13	0.06	-0.48	11.00	False	False
’covariance type’: ’spherical’, ’n ’: 2	0.50	0.39	0.36	0.09	-0.32	2.00	True	True
’covariance type’: ’spherical’, ’n ’: 3	0.57	0.57	0.20	0.05	-0.55	3.00	False	False
’covariance type’: ’spherical’, ’n ’: 4	0.54	0.54	0.23	0.07	-0.54	4.00	False	False
’covariance type’: ’spherical’, ’n ’: 5	0.54	0.61	0.26	0.07	-0.35	5.00	False	False
’covariance type’: ’spherical’, ’n ’: 6	0.59	0.54	0.18	0.07	-0.37	6.00	False	False
’covariance type’: ’spherical’, ’n ’: 7	0.57	0.59	0.31	0.09	-0.37	7.00	True	True
’covariance type’: ’spherical’, ’n ’: 8	0.57	0.60	0.31	0.08	-0.37	8.00	False	False
’covariance type’: ’spherical’, ’n ’: 9	0.63	0.61	0.24	0.06	-0.39	9.00	False	False
’covariance type’: ’spherical’, ’n ’: 10	0.60	0.60	0.22	0.08	-0.39	10.00	False	False
’covariance type’: ’spherical’, ’n ’: 11	0.62	0.59	0.23	0.09	-0.35	11.00	False	False

Table C.104: TEP dataset - GMM clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’covariance type’: ’full’, ’n ’: 11	0.51	0.51	0.62	0.17	-0.70	11.00	False	False
’covariance type’: ’tied’, ’n ’: 2	0.15	0.15	0.72	0.14	-0.58	2.00	True	True
’covariance type’: ’tied’, ’n ’: 3	0.18	0.18	0.72	0.13	-0.70	3.00	True	True
’covariance type’: ’tied’, ’n ’: 4	0.38	0.38	0.60	0.14	-0.80	4.00	False	False
’covariance type’: ’tied’, ’n ’: 5	0.31	0.31	0.57	0.15	-0.78	5.00	False	False
’covariance type’: ’tied’, ’n ’: 6	0.35	0.35	0.62	0.15	0.08	6.00	True	True
’covariance type’: ’tied’, ’n ’: 7	0.50	0.50	0.69	0.17	-0.65	7.00	False	False
’covariance type’: ’tied’, ’n ’: 8	0.49	0.49	0.70	0.18	-0.66	8.00	True	True
’covariance type’: ’tied’, ’n ’: 9	0.49	0.50	0.69	0.19	-0.64	9.00	True	True
’covariance type’: ’tied’, ’n ’: 10	0.49	0.50	0.69	0.19	-0.69	10.00	False	False
’covariance type’: ’tied’, ’n ’: 11	0.49	0.49	0.69	0.18	-0.69	10.00	False	False
’covariance type’: ’diag’, ’n ’: 2	0.51	0.51	0.37	0.02	-0.38	2.00	False	False
’covariance type’: ’diag’, ’n ’: 3	0.54	0.54	0.38	0.10	-0.35	3.00	False	False
’covariance type’: ’diag’, ’n ’: 4	0.59	0.59	0.59	0.08	-0.09	4.00	False	False
’covariance type’: ’diag’, ’n ’: 5	0.52	0.57	0.57	0.14	-0.33	5.00	False	False
’covariance type’: ’diag’, ’n ’: 6	0.56	0.56	0.66	0.17	-0.09	6.00	True	True
’covariance type’: ’diag’, ’n ’: 7	0.55	0.55	0.59	0.17	-0.05	7.00	False	False
’covariance type’: ’diag’, ’n ’: 8	0.54	0.54	0.67	0.15	-0.41	8.00	False	False
’covariance type’: ’diag’, ’n ’: 9	0.53	0.54	0.60	0.18	-0.38	9.00	False	False
’covariance type’: ’diag’, ’n ’: 10	0.53	0.52	0.68	0.16	-0.34	10.00	False	False
’covariance type’: ’diag’, ’n ’: 11	0.53	0.52	0.62	0.18	-0.04	11.00	True	True
’covariance type’: ’spherical’, ’n ’: 2	0.51	0.51	0.37	0.02	-0.38	2.00	False	False
’covariance type’: ’spherical’, ’n ’: 3	0.63	0.51	0.41	0.03	-0.36	3.00	False	False
’covariance type’: ’spherical’, ’n ’: 4	0.61	0.54	0.56	0.12	-0.72	4.00	False	False
’covariance type’: ’spherical’, ’n ’: 5	0.37	0.58	0.65	0.14	-0.35	5.00	False	False
’covariance type’: ’spherical’, ’n ’: 6	0.56	0.56	0.67	0.16	-0.12	6.00	False	False
’covariance type’: ’spherical’, ’n ’: 7	0.55	0.55	0.67	0.17	-0.10	7.00	True	True
’covariance type’: ’spherical’, ’n ’: 8	0.54	0.54	0.69	0.19	-0.32	8.00	True	True
’covariance type’: ’spherical’, ’n ’: 9	0.55	0.54	0.68	0.18	-0.13	9.00	True	True
’covariance type’: ’spherical’, ’n ’: 10	0.53	0.55	0.67	0.18	-0.30	10.00	False	False
’covariance type’: ’spherical’, ’n ’: 11	0.53	0.53	0.62	0.17	-0.62	11.00	False	False

Table C.105: TEP dataset - GMM clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’covariance type’: ’full’, ’n ’: 2	0.48	0.48	0.43	0.12	-0.57	2.00	False	False
’covariance type’: ’full’, ’n ’: 3	0.53	0.53	0.40	0.08	-0.69	3.00	False	False
’covariance type’: ’full’, ’n ’: 4	0.50	0.50	0.44	0.09	-0.73	4.00	False	False
’covariance type’: ’full’, ’n ’: 5	0.57	0.57	0.48	0.14	-0.41	5.00	True	True
’covariance type’: ’tied’, ’n ’: 2	0.33	0.15	0.57	0.09	-0.46	2.00	True	True
’covariance type’: ’tied’, ’n ’: 3	0.32	0.32	0.38	0.10	-0.77	3.00	False	False
’covariance type’: ’tied’, ’n ’: 4	0.43	0.43	0.23	0.12	-0.76	4.00	False	False
’covariance type’: ’tied’, ’n ’: 5	0.44	0.44	0.46	0.14	-0.80	5.00	False	False
’covariance type’: ’tied’, ’n ’: 6	0.51	0.51	0.48	0.15	-0.75	6.00	True	True
’covariance type’: ’tied’, ’n ’: 7	0.46	0.56	0.49	0.14	-0.76	7.00	True	True
’covariance type’: ’tied’, ’n ’: 8	0.48	0.56	0.41	0.14	-0.70	8.00	False	False
’covariance type’: ’tied’, ’n ’: 9	0.53	0.46	0.38	0.15	-0.70	9.00	True	True
’covariance type’: ’tied’, ’n ’: 10	0.48	0.48	0.37	0.15	-0.61	10.00	False	False
’covariance type’: ’tied’, ’n ’: 11	0.48	0.49	0.39	0.15	-0.71	11.00	True	True
’covariance type’: ’diag’, ’n ’: 2	0.20	0.21	0.43	0.09	-0.86	2.00	False	False
’covariance type’: ’diag’, ’n ’: 3	0.58	0.58	0.39	0.06	-0.73	3.00	False	False
’covariance type’: ’diag’, ’n ’: 4	0.50	0.50	0.45	0.08	-0.49	4.00	False	False
’covariance type’: ’diag’, ’n ’: 5	0.58	0.58	0.47	0.13	-0.50	5.00	False	False
’covariance type’: ’diag’, ’n ’: 6	0.60	0.60	0.48	0.11	-0.57	6.00	True	True
’covariance type’: ’diag’, ’n ’: 7	0.50	0.50	0.45	0.14	-0.69	7.00	True	True
’covariance type’: ’diag’, ’n ’: 8	0.49	0.49	0.41	0.14	-0.71	8.00	False	False
’covariance type’: ’diag’, ’n ’: 9	0.49	0.48	0.39	0.14	-0.75	9.00	False	False
’covariance type’: ’diag’, ’n ’: 10	0.47	0.51	0.36	0.14	-0.71	10.00	False	False
’covariance type’: ’diag’, ’n ’: 11	0.50	0.50	0.37	0.13	-0.68	11.00	False	False
’covariance type’: ’spherical’, ’n ’: 2	0.48	0.48	0.43	0.02	-0.42	2.00	False	False
’covariance type’: ’spherical’, ’n ’: 3	0.51	0.51	0.33	0.06	-0.74	3.00	False	False
’covariance type’: ’spherical’, ’n ’: 4	0.57	0.57	0.39	0.07	-0.69	4.00	False	False
’covariance type’: ’spherical’, ’n ’: 5	0.56	0.46	0.44	0.12	-0.69	5.00	False	False
’covariance type’: ’spherical’, ’n ’: 6	0.54	0.55	0.44	0.12	-0.74	6.00	False	False
’covariance type’: ’spherical’, ’n ’: 7	0.54	0.54	0.42	0.14	-0.72	7.00	False	False
’covariance type’: ’spherical’, ’n ’: 8	0.53	0.53	0.43	0.14	-0.69	8.00	True	True
’covariance type’: ’spherical’, ’n ’: 9	0.51	0.52	0.41	0.15	-0.59	9.00	True	True
’covariance type’: ’spherical’, ’n ’: 10	0.51	0.49	0.43	0.15	-0.68	10.00	True	True
’covariance type’: ’spherical’, ’n ’: 11	0.50	0.47	0.38	0.14	-0.65	11.00	False	False

Table C.106: TEP dataset - GMM clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'covariance type': 'full', 'n ': 4	0.69	0.69	0.67	0.16	-0.42	4.00	False	False
'covariance type': 'full', 'n ': 5	0.67	0.44	0.69	0.17	-0.11	5.00	True	True
'covariance type': 'full', 'n ': 6	0.65	0.42	0.66	0.18	-0.17	6.00	True	True
'covariance type': 'full', 'n ': 7	0.63	0.64	0.66	0.18	-0.22	7.00	True	True
'covariance type': 'full', 'n ': 8	0.62	0.62	0.67	0.17	-0.26	8.00	False	False
'covariance type': 'full', 'n ': 9	0.61	0.61	0.66	0.17	-0.45	9.00	False	False
'covariance type': 'full', 'n ': 10	0.61	0.60	0.64	0.17	-0.45	10.00	False	False
'covariance type': 'full', 'n ': 11	0.59	0.59	0.65	0.19	-0.44	11.00	False	False
'covariance type': 'tied', 'n ': 2	0.17	0.17	0.74	0.16	-0.79	2.00	True	True
'covariance type': 'tied', 'n ': 3	0.20	0.25	0.70	0.16	-0.64	3.00	True	True
'covariance type': 'tied', 'n ': 4	0.25	0.26	0.71	0.15	-0.77	4.00	True	True
'covariance type': 'tied', 'n ': 5	0.53	0.53	0.65	0.17	-0.57	4.00	False	False
'covariance type': 'tied', 'n ': 6	0.59	0.59	0.67	0.20	-0.60	6.00	False	False
'covariance type': 'tied', 'n ': 7	0.57	0.58	0.68	0.20	-0.62	7.00	False	False
'covariance type': 'tied', 'n ': 8	0.57	0.57	0.69	0.20	-0.60	8.00	True	True
'covariance type': 'tied', 'n ': 9	0.56	0.55	0.69	0.19	-0.48	9.00	True	True
'covariance type': 'tied', 'n ': 10	0.56	0.56	0.69	0.18	-0.44	10.00	True	True
'covariance type': 'tied', 'n ': 11	0.56	0.56	0.66	0.19	-0.48	11.00	False	False
'covariance type': 'diag', 'n ': 2	0.48	0.48	0.64	0.08	-0.47	2.00	False	False
'covariance type': 'diag', 'n ': 3	0.44	0.44	0.70	0.13	-0.33	3.00	True	True
'covariance type': 'diag', 'n ': 4	0.67	0.41	0.61	0.14	-0.34	4.00	False	False
'covariance type': 'diag', 'n ': 5	0.65	0.65	0.62	0.17	-0.63	5.00	False	False
'covariance type': 'diag', 'n ': 6	0.64	0.64	0.64	0.19	-0.23	6.00	False	False
'covariance type': 'diag', 'n ': 7	0.63	0.62	0.65	0.17	-0.23	7.00	False	False
'covariance type': 'diag', 'n ': 8	0.62	0.62	0.65	0.19	-0.25	8.00	False	False
'covariance type': 'diag', 'n ': 9	0.61	0.61	0.64	0.18	-0.19	9.00	False	False
'covariance type': 'diag', 'n ': 10	0.60	0.60	0.66	0.17	-0.47	10.00	False	False
'covariance type': 'diag', 'n ': 11	0.59	0.59	0.65	0.17	-0.47	11.00	False	False
'covariance type': 'spherical', 'n ': 2	0.48	0.48	0.64	0.08	-0.43	2.00	False	False
'covariance type': 'spherical', 'n ': 3	0.44	0.44	0.65	0.12	-0.51	3.00	False	False
'covariance type': 'spherical', 'n ': 4	0.68	0.68	0.60	0.12	-0.54	4.00	False	False
'covariance type': 'spherical', 'n ': 5	0.65	0.65	0.64	0.15	-0.30	5.00	False	False
'covariance type': 'spherical', 'n ': 6	0.40	0.64	0.67	0.19	-0.56	6.00	False	False
'covariance type': 'spherical', 'n ': 7	0.62	0.63	0.67	0.19	-0.24	7.00	False	False
'covariance type': 'spherical', 'n ': 8	0.61	0.62	0.68	0.20	-0.24	8.00	True	True
'covariance type': 'spherical', 'n ': 9	0.61	0.61	0.67	0.20	-0.23	9.00	True	True
'covariance type': 'spherical', 'n ': 10	0.59	0.59	0.67	0.18	-0.41	10.00	False	False
'covariance type': 'spherical', 'n ': 11	0.59	0.58	0.65	0.18	-0.45	11.00	False	False

Table C.107: TEP dataset - GMM clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’covariance type’: ’full’, ’n ’: 2	0.45	0.47	0.43	0.14	-0.76	2.00	False	False
’covariance type’: ’full’, ’n ’: 3	0.65	0.65	0.46	0.12	-0.74	3.00	False	False
’covariance type’: ’full’, ’n ’: 4	0.62	0.62	0.49	0.14	-0.79	4.00	False	False
’covariance type’: ’full’, ’n ’: 5	0.58	0.58	0.40	0.14	-0.88	5.00	False	False
’covariance type’: ’full’, ’n ’: 6	0.61	0.69	0.40	0.13	-0.79	6.00	False	False
’covariance type’: ’full’, ’n ’: 7	0.63	0.59	0.45	0.12	-0.71	7.00	False	False
’covariance type’: ’full’, ’n ’: 8	0.69	0.69	0.49	0.17	-0.57	8.00	True	True
’covariance type’: ’full’, ’n ’: 9	0.64	0.64	0.47	0.13	-0.64	9.00	False	False
’covariance type’: ’full’, ’n ’: 10	0.67	0.63	0.46	0.14	-0.71	10.00	False	False
’covariance type’: ’full’, ’n ’: 11	0.64	0.64	0.50	0.15	-0.58	11.00	True	True
’covariance type’: ’tied’, ’n ’: 2	0.37	0.41	0.38	0.10	-0.83	2.00	False	False
’covariance type’: ’tied’, ’n ’: 3	0.53	0.53	0.43	0.11	-0.82	3.00	False	False
’covariance type’: ’tied’, ’n ’: 4	0.64	0.55	0.50	0.14	-0.74	4.00	False	False
’covariance type’: ’tied’, ’n ’: 5	0.69	0.69	0.47	0.11	-0.72	5.00	False	False
’covariance type’: ’tied’, ’n ’: 6	0.64	0.66	0.46	0.15	-0.82	6.00	False	False
’covariance type’: ’tied’, ’n ’: 7	0.74	0.74	0.49	0.15	-0.78	7.00	False	False
’covariance type’: ’tied’, ’n ’: 8	0.73	0.73	0.51	0.16	-0.64	8.00	True	True
’covariance type’: ’tied’, ’n ’: 9	0.68	0.67	0.51	0.14	-0.73	9.00	True	True
’covariance type’: ’tied’, ’n ’: 10	0.66	0.66	0.50	0.15	-0.64	10.00	True	True
’covariance type’: ’tied’, ’n ’: 11	0.65	0.64	0.48	0.13	-0.75	11.00	False	False
’covariance type’: ’diag’, ’n ’: 2	0.36	0.35	0.38	0.15	-0.93	2.00	False	False
’covariance type’: ’diag’, ’n ’: 3	0.22	0.61	0.45	0.12	-0.77	3.00	False	False
’covariance type’: ’diag’, ’n ’: 4	0.64	0.64	0.48	0.12	-0.68	4.00	False	False
’covariance type’: ’diag’, ’n ’: 5	0.64	0.64	0.46	0.13	-0.72	5.00	False	False
’covariance type’: ’diag’, ’n ’: 6	0.77	0.77	0.45	0.14	-0.76	6.00	False	False
’covariance type’: ’diag’, ’n ’: 7	0.71	0.74	0.41	0.15	-0.59	7.00	False	False
’covariance type’: ’diag’, ’n ’: 8	0.69	0.67	0.44	0.16	-0.55	8.00	True	True
’covariance type’: ’diag’, ’n ’: 9	0.67	0.67	0.46	0.13	-0.61	9.00	False	False
’covariance type’: ’diag’, ’n ’: 10	0.65	0.65	0.48	0.14	-0.72	10.00	False	False
’covariance type’: ’diag’, ’n ’: 11	0.64	0.63	0.46	0.13	-0.67	11.00	False	False
’covariance type’: ’spherical’, ’n ’: 2	0.27	0.27	0.38	0.16	-0.89	2.00	False	False
’covariance type’: ’spherical’, ’n ’: 3	0.48	0.48	0.41	0.11	-0.81	3.00	False	False
’covariance type’: ’spherical’, ’n ’: 4	0.58	0.64	0.48	0.13	-0.78	4.00	False	False
’covariance type’: ’spherical’, ’n ’: 5	0.59	0.65	0.46	0.14	-0.67	5.00	False	False
’covariance type’: ’spherical’, ’n ’: 6	0.64	0.64	0.46	0.15	-0.76	6.00	False	False
’covariance type’: ’spherical’, ’n ’: 7	0.71	0.71	0.44	0.15	-0.65	7.00	False	False
’covariance type’: ’spherical’, ’n ’: 8	0.70	0.70	0.49	0.17	-0.59	8.00	True	True
’covariance type’: ’spherical’, ’n ’: 9	0.68	0.69	0.50	0.14	-0.51	9.00	True	True
’covariance type’: ’spherical’, ’n ’: 10	0.68	0.66	0.50	0.15	-0.61	10.00	True	True
’covariance type’: ’spherical’, ’n ’: 11	0.65	0.65	0.45	0.15	-0.60	11.00	False	False

Table C.108: TEP dataset - GMM clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’covariance type’: ’full’, ’n ’: 2	0.30	0.30	0.47	0.09	-0.83	2.00	False	False
’covariance type’: ’full’, ’n ’: 3	0.60	0.60	0.56	0.14	-0.67	3.00	False	False
’covariance type’: ’full’, ’n ’: 4	0.63	0.63	0.60	0.18	-0.63	4.00	True	True
’covariance type’: ’full’, ’n ’: 5	0.62	0.62	0.50	0.15	-0.62	5.00	False	False
’covariance type’: ’full’, ’n ’: 6	0.54	0.63	0.46	0.14	-0.67	6.00	False	False
’covariance type’: ’full’, ’n ’: 7	0.62	0.62	0.48	0.15	-0.67	7.00	False	False
’covariance type’: ’full’, ’n ’: 8	0.61	0.61	0.44	0.15	-0.60	8.00	False	False
’covariance type’: ’full’, ’n ’: 9	0.62	0.61	0.45	0.14	-0.66	9.00	False	False
’covariance type’: ’full’, ’n ’: 10	0.59	0.60	0.46	0.15	-0.57	10.00	False	False
’covariance type’: ’full’, ’n ’: 11	0.58	0.59	0.41	0.13	-0.60	11.00	False	False
’covariance type’: ’tied’, ’n ’: 2	0.34	0.37	0.49	0.10	-0.82	2.00	False	False
’covariance type’: ’tied’, ’n ’: 3	0.58	0.58	0.56	0.14	-0.50	3.00	True	True
’covariance type’: ’tied’, ’n ’: 4	0.63	0.63	0.61	0.18	-0.67	4.00	True	True
’covariance type’: ’tied’, ’n ’: 5	0.63	0.62	0.59	0.19	-0.60	5.00	True	True
’covariance type’: ’tied’, ’n ’: 6	0.63	0.63	0.41	0.18	-0.67	6.00	False	False
’covariance type’: ’tied’, ’n ’: 7	0.61	0.62	0.45	0.16	-0.67	7.00	False	False
’covariance type’: ’tied’, ’n ’: 8	0.61	0.62	0.47	0.16	-0.65	8.00	False	False
’covariance type’: ’tied’, ’n ’: 9	0.62	0.59	0.47	0.15	-0.62	9.00	False	False
’covariance type’: ’tied’, ’n ’: 10	0.59	0.61	0.48	0.16	-0.63	10.00	False	False
’covariance type’: ’tied’, ’n ’: 11	0.57	0.57	0.45	0.14	-0.65	11.00	False	False
’covariance type’: ’diag’, ’n ’: 2	0.32	0.32	0.46	0.11	-0.74	2.00	False	False
’covariance type’: ’diag’, ’n ’: 3	0.58	0.58	0.56	0.14	-0.50	3.00	False	False
’covariance type’: ’diag’, ’n ’: 4	0.65	0.65	0.59	0.17	-0.66	4.00	False	False
’covariance type’: ’diag’, ’n ’: 5	0.63	0.63	0.51	0.14	-0.59	5.00	False	False
’covariance type’: ’diag’, ’n ’: 6	0.61	0.54	0.47	0.10	-0.58	6.00	False	False
’covariance type’: ’diag’, ’n ’: 7	0.67	0.66	0.49	0.12	-0.54	7.00	False	False
’covariance type’: ’diag’, ’n ’: 8	0.64	0.52	0.47	0.15	-0.55	8.00	True	True
’covariance type’: ’diag’, ’n ’: 9	0.62	0.63	0.46	0.15	-0.57	9.00	True	True
’covariance type’: ’diag’, ’n ’: 10	0.61	0.59	0.41	0.13	-0.62	10.00	False	False
’covariance type’: ’diag’, ’n ’: 11	0.60	0.59	0.41	0.13	-0.59	11.00	False	False
’covariance type’: ’spherical’, ’n ’: 2	0.36	0.36	0.47	0.09	-0.67	2.00	False	False
’covariance type’: ’spherical’, ’n ’: 3	0.64	0.64	0.53	0.10	-0.80	3.00	False	False
’covariance type’: ’spherical’, ’n ’: 4	0.63	0.63	0.60	0.16	-0.65	4.00	True	True
’covariance type’: ’spherical’, ’n ’: 5	0.62	0.62	0.58	0.19	-0.66	5.00	False	False
’covariance type’: ’spherical’, ’n ’: 6	0.62	0.61	0.48	0.17	-0.70	6.00	False	False
’covariance type’: ’spherical’, ’n ’: 7	0.60	0.58	0.47	0.17	-0.61	7.00	False	False
’covariance type’: ’spherical’, ’n ’: 8	0.63	0.63	0.43	0.14	-0.64	8.00	False	False
’covariance type’: ’spherical’, ’n ’: 9	0.62	0.62	0.45	0.16	-0.59	9.00	True	True
’covariance type’: ’spherical’, ’n ’: 10	0.60	0.61	0.46	0.14	-0.51	10.00	False	False
’covariance type’: ’spherical’, ’n ’: 11	0.59	0.59	0.44	0.15	-0.53	11.00	True	True

Table C.109: TEP dataset - DBSCAN clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’eps’: 4.4, ’min samples’: 1200	0.47	0.47	0.18	0.02	nan	2.00	False	False
’eps’: 4.4, ’min samples’: 1400	0.47	0.47	0.18	0.02	nan	2.00	False	False
’eps’: 4.4, ’min samples’: 1600	0.46	0.46	0.17	0.02	nan	2.00	False	False
’eps’: 4.4, ’min samples’: 1800	0.46	0.46	0.17	0.02	nan	2.00	False	False
’eps’: 4.6, ’min samples’: 200	0.50	0.50	0.36	0.03	nan	2.00	True	False
’eps’: 4.6, ’min samples’: 400	0.50	0.50	0.36	0.03	nan	2.00	False	False
’eps’: 4.6, ’min samples’: 600	0.50	0.50	0.36	0.03	nan	2.00	False	False
’eps’: 4.6, ’min samples’: 800	0.50	0.50	0.36	0.03	nan	2.00	False	False
’eps’: 4.6, ’min samples’: 1000	0.64	0.64	0.18	0.03	-0.29	3.00	True	True
’eps’: 4.6, ’min samples’: 1200	0.63	0.63	0.18	0.03	-0.30	3.00	False	False
’eps’: 4.6, ’min samples’: 1400	0.46	0.46	0.19	0.02	nan	2.00	False	False
’eps’: 4.6, ’min samples’: 1600	0.47	0.47	0.19	0.02	nan	2.00	False	False
’eps’: 4.6, ’min samples’: 1800	0.48	0.48	0.18	0.02	nan	2.00	False	False

Table C.110: TEP dataset - DBSCAN clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'eps': 0.4, 'min samples': 200	0.54	0.54	0.44	0.02	0.56	3.00	True	True
'eps': 0.4, 'min samples': 400	0.54	0.54	0.40	0.02	0.54	3.00	False	False
'eps': 0.4, 'min samples': 600	0.51	0.51	0.34	0.02	nan	2.00	False	False
'eps': 0.4, 'min samples': 800	0.50	0.50	0.32	0.02	nan	2.00	False	False
'eps': 0.4, 'min samples': 1000	0.48	0.48	0.31	0.02	nan	2.00	False	False
'eps': 0.4, 'min samples': 1200	0.46	0.46	0.29	0.02	nan	2.00	False	False
'eps': 0.4, 'min samples': 1400	0.43	0.43	0.25	0.02	nan	2.00	False	False
'eps': 0.4, 'min samples': 1600	0.39	0.39	0.21	0.02	nan	2.00	False	False
'eps': 0.4, 'min samples': 1800	0.32	0.32	0.12	0.02	nan	2.00	False	False
'eps': 0.6, 'min samples': 200	0.63	0.63	0.53	0.03	0.11	3.00	True	True
'eps': 0.6, 'min samples': 400	0.55	0.55	0.48	0.03	0.51	3.00	True	True
'eps': 0.6, 'min samples': 600	0.54	0.54	0.46	0.02	0.54	3.00	True	True
'eps': 0.6, 'min samples': 800	0.51	0.51	0.37	0.02	nan	2.00	False	False
'eps': 0.6, 'min samples': 1000	0.51	0.51	0.37	0.02	nan	2.00	False	False
'eps': 0.6, 'min samples': 1200	0.51	0.51	0.37	0.02	nan	2.00	False	False
'eps': 0.6, 'min samples': 1400	0.51	0.51	0.37	0.02	nan	2.00	False	False
'eps': 0.6, 'min samples': 1600	0.51	0.51	0.36	0.02	nan	2.00	False	False
'eps': 0.6, 'min samples': 1800	0.52	0.52	0.36	0.02	nan	2.00	False	False
'eps': 0.8, 'min samples': 200	0.47	0.47	0.57	0.03	nan	2.00	False	False
'eps': 0.8, 'min samples': 400	0.61	0.61	0.53	0.03	-0.13	3.00	False	False
'eps': 0.8, 'min samples': 600	0.55	0.55	0.50	0.03	0.39	3.00	True	True
'eps': 0.8, 'min samples': 800	0.54	0.54	0.49	0.03	0.41	3.00	True	True
'eps': 0.8, 'min samples': 1000	0.49	0.49	0.39	0.02	nan	2.00	False	False
'eps': 0.8, 'min samples': 1200	0.50	0.50	0.38	0.02	nan	2.00	False	False
'eps': 0.8, 'min samples': 1400	0.50	0.50	0.38	0.02	nan	2.00	False	False
'eps': 0.8, 'min samples': 1600	0.50	0.50	0.38	0.02	nan	2.00	False	False
'eps': 0.8, 'min samples': 1800	0.50	0.50	0.38	0.02	nan	2.00	False	False
'eps': 1, 'min samples': 200	0.46	0.46	0.58	0.03	nan	2.00	True	False
'eps': 1, 'min samples': 400	0.46	0.46	0.57	0.03	nan	2.00	False	False
'eps': 1, 'min samples': 600	0.60	0.60	0.54	0.03	-0.16	3.00	True	True
'eps': 1, 'min samples': 800	0.57	0.57	0.53	0.03	0.06	3.00	True	True
'eps': 1, 'min samples': 1000	0.56	0.56	0.51	0.03	0.19	3.00	True	True
'eps': 1, 'min samples': 1200	0.48	0.48	0.39	0.02	nan	2.00	False	False

Table C.111: TEP dataset - DBSCAN clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’eps’: 0.8, ’min samples’: 50	0.61	0.61	0.27	0.06	-0.24	10.00	True	True
’eps’: 0.8, ’min samples’: 100	0.65	0.65	0.31	0.01	-0.13	3.00	False	False
’eps’: 0.8, ’min samples’: 200	0.43	0.43	0.10	0.02	0.29	3.00	True	True
’eps’: 0.8, ’min samples’: 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 800	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 1000	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 1200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 0.8, ’min samples’: 1400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1, ’min samples’: 50	0.44	0.44	0.37	0.06	-0.04	5.00	True	True
’eps’: 1, ’min samples’: 100	0.68	0.68	0.32	0.02	-0.14	3.00	False	False
’eps’: 1, ’min samples’: 200	0.62	0.62	0.30	0.01	-0.13	3.00	False	False
’eps’: 1, ’min samples’: 400	0.35	0.35	-0.02	0.02	0.26	3.00	False	False
’eps’: 1, ’min samples’: 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1, ’min samples’: 800	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1, ’min samples’: 1000	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1, ’min samples’: 1200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1, ’min samples’: 1400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1.5, ’min samples’: 50	0.43	0.43	0.41	0.03	-0.22	6.00	True	True
’eps’: 1.5, ’min samples’: 100	0.43	0.43	0.27	0.06	-0.08	7.00	False	False
’eps’: 1.5, ’min samples’: 200	0.50	0.50	0.43	0.02	nan	2.00	False	False
’eps’: 1.5, ’min samples’: 400	0.67	0.67	0.32	0.02	-0.25	3.00	False	False
’eps’: 1.5, ’min samples’: 600	0.48	0.48	0.23	0.02	0.36	3.00	True	True
’eps’: 1.5, ’min samples’: 800	0.34	0.34	0.08	0.02	nan	2.00	False	False
’eps’: 1.5, ’min samples’: 1000	0.28	0.28	-0.05	0.01	nan	2.00	False	False
’eps’: 1.5, ’min samples’: 1200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 1.5, ’min samples’: 1400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’eps’: 2, ’min samples’: 50	0.00	0.00	0.34	0.04	nan	2.00	True	False
’eps’: 2, ’min samples’: 100	0.21	0.21	0.42	0.05	-0.55	4.00	True	True
’eps’: 2, ’min samples’: 200	0.49	0.49	0.43	0.02	nan	2.00	True	False

Table C.112: TEP dataset - DBSCAN clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
eps’: 1, ’min samples’: 50	0.45	0.45	0.49	0.11	0.59	8.00	True	True
’eps’: 1, ’min samples’: 100	0.70	0.70	0.53	0.09	-0.28	3.00	False	False
’eps’: 1, ’min samples’: 200	0.63	0.63	0.43	0.08	0.32	4.00	False	False
’eps’: 1, ’min samples’: 400	0.58	0.58	0.42	0.07	0.54	3.00	False	False
’eps’: 1, ’min samples’: 600	0.55	0.55	0.38	0.07	0.52	3.00	False	False
’eps’: 1, ’min samples’: 800	0.49	0.49	0.27	0.05	nan	2.00	False	False
’eps’: 1, ’min samples’: 1000	0.45	0.45	0.23	0.04	nan	2.00	False	False
’eps’: 1, ’min samples’: 1200	0.40	0.40	0.18	0.04	nan	2.00	False	False
’eps’: 1, ’min samples’: 1400	0.37	0.37	0.14	0.04	nan	2.00	False	False
’eps’: 1.5, ’min samples’: 50	0.43	0.43	0.58	0.12	0.55	7.00	True	True
’eps’: 1.5, ’min samples’: 100	0.48	0.48	0.57	0.11	0.62	4.00	True	True
’eps’: 1.5, ’min samples’: 200	0.71	0.71	0.54	0.09	-0.28	3.00	False	False
’eps’: 1.5, ’min samples’: 400	0.61	0.61	0.49	0.08	0.46	3.00	False	False
’eps’: 1.5, ’min samples’: 600	0.60	0.60	0.46	0.07	0.54	3.00	False	False
’eps’: 1.5, ’min samples’: 800	0.59	0.59	0.44	0.07	0.54	3.00	False	False
’eps’: 1.5, ’min samples’: 1000	0.56	0.56	0.33	0.05	nan	2.00	False	False
’eps’: 1.5, ’min samples’: 1200	0.56	0.56	0.32	0.05	nan	2.00	False	False
’eps’: 1.5, ’min samples’: 1400	0.56	0.56	0.32	0.05	nan	2.00	False	False
’eps’: 2, ’min samples’: 50	0.41	0.41	0.48	0.10	0.36	9.00	False	False
’eps’: 2, ’min samples’: 100	0.46	0.46	0.58	0.12	0.55	4.00	True	True
’eps’: 2, ’min samples’: 200	0.50	0.50	0.63	0.08	nan	2.00	False	False
’eps’: 2, ’min samples’: 400	0.71	0.71	0.54	0.09	-0.30	3.00	False	False
’eps’: 2, ’min samples’: 600	0.60	0.60	0.50	0.08	0.42	3.00	False	False
’eps’: 2, ’min samples’: 800	0.60	0.60	0.49	0.08	0.45	3.00	False	False
’eps’: 2, ’min samples’: 1000	0.59	0.59	0.47	0.07	0.48	3.00	False	False
’eps’: 2, ’min samples’: 1200	0.54	0.54	0.34	0.05	nan	2.00	False	False
’eps’: 2, ’min samples’: 1400	0.54	0.54	0.34	0.05	nan	2.00	False	False
’eps’: 2.5, ’min samples’: 50	0.32	0.32	0.64	0.08	-0.27	8.00	True	True
’eps’: 2.5, ’min samples’: 100	0.43	0.43	0.59	0.12	0.54	6.00	True	True
’eps’: 2.5, ’min samples’: 200	0.49	0.49	0.64	0.08	nan	2.00	True	False

Table C.113: TEP dataset - DBSCAN clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'eps': 3, 'min samples': 10	0.76	0.76	-0.04	0.12	-0.29	7.00	True	True
'eps': 3, 'min samples': 20	0.76	0.76	0.16	0.06	-0.28	6.00	False	False
'eps': 3, 'min samples': 50	0.72	0.72	0.28	0.04	0.04	9.00	False	False
'eps': 3, 'min samples': 100	0.33	0.33	-0.22	0.08	0.19	15.00	True	True
'eps': 3, 'min samples': 150	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 3.5, 'min samples': 10	0.77	0.77	0.45	0.10	-0.23	3.00	True	True
'eps': 3.5, 'min samples': 20	0.77	0.77	0.22	0.07	-0.23	4.00	False	False
'eps': 3.5, 'min samples': 50	0.75	0.75	0.30	0.03	-0.14	8.00	False	False
'eps': 3.5, 'min samples': 100	0.52	0.53	0.22	0.08	-0.03	18.00	False	False
'eps': 3.5, 'min samples': 150	0.13	0.13	-0.32	0.04	0.09	4.00	False	False
'eps': 3.5, 'min samples': 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 3.5, 'min samples': 250	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 3.5, 'min samples': 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 4, 'min samples': 10	0.53	0.53	0.38	0.09	0.10	2.00	True	True
'eps': 4, 'min samples': 20	0.53	0.53	0.38	0.09	0.10	2.00	False	False
'eps': 4, 'min samples': 50	0.75	0.75	0.16	0.06	-0.45	5.00	False	False
'eps': 4, 'min samples': 100	0.66	0.66	0.41	0.08	-0.13	10.00	True	True
'eps': 4, 'min samples': 150	0.39	0.39	-0.08	0.05	0.13	7.00	True	True
'eps': 4.5, 'min samples': 10	0.53	0.53	0.38	0.09	0.10	2.00	False	False
'eps': 4.5, 'min samples': 20	0.53	0.53	0.38	0.09	0.10	2.00	False	False
'eps': 4.5, 'min samples': 50	0.77	0.77	0.26	0.05	-0.43	4.00	False	False
'eps': 4.5, 'min samples': 100	0.74	0.74	0.34	0.06	-0.19	7.00	False	False
'eps': 4.5, 'min samples': 150	0.56	0.56	0.21	0.02	0.11	9.00	True	True
'eps': 4.5, 'min samples': 200	0.34	0.34	-0.12	0.05	0.22	5.00	True	True
'eps': 4.5, 'min samples': 250	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 4.5, 'min samples': 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 5, 'min samples': 150	0.62	0.62	0.39	0.09	-0.26	12.00	False	False
'eps': 5, 'min samples': 200	0.53	0.53	0.13	0.03	0.20	6.00	True	True
'eps': 5, 'min samples': 250	0.34	0.34	-0.11	0.04	0.24	5.00	True	True
'eps': 5, 'min samples': 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False

Table C.114: TEP dataset - DBSCAN clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'eps': 3, 'min samples': 100	0.44	0.44	0.35	0.09	0.32	6.00	True	True
'eps': 3, 'min samples': 200	0.70	0.70	0.52	0.09	0.26	3.00	True	True
'eps': 3, 'min samples': 300	0.56	0.56	0.19	0.04	-0.13	5.00	False	False
'eps': 3, 'min samples': 400	0.38	0.38	0.05	0.07	0.32	3.00	False	False
'eps': 3, 'min samples': 500	0.17	0.17	-0.05	0.05	nan	2.00	False	False
'eps': 3, 'min samples': 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 3, 'min samples': 800	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 3, 'min samples': 1000	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 3.5, 'min samples': 100	0.43	0.43	0.36	0.10	0.12	7.00	True	True
'eps': 3.5, 'min samples': 200	0.72	0.72	0.53	0.10	-0.29	3.00	True	True
'eps': 3.5, 'min samples': 300	0.63	0.63	0.40	0.09	0.20	4.00	True	True
'eps': 3.5, 'min samples': 400	0.56	0.56	0.19	0.04	-0.02	5.00	False	False
'eps': 3.5, 'min samples': 500	0.45	0.45	0.18	0.07	0.38	3.00	False	False
'eps': 3.5, 'min samples': 600	0.32	0.32	0.08	0.05	nan	2.00	False	False
'eps': 3.5, 'min samples': 800	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 3.5, 'min samples': 1000	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 4, 'min samples': 100	0.46	0.46	0.31	0.15	0.06	4.00	True	True
'eps': 4, 'min samples': 200	0.51	0.51	0.47	0.08	nan	2.00	False	False
'eps': 4, 'min samples': 300	0.71	0.71	0.53	0.10	-0.21	3.00	True	True
'eps': 4, 'min samples': 400	0.61	0.61	0.45	0.07	0.49	3.00	True	True
'eps': 4, 'min samples': 500	0.57	0.57	0.19	0.05	-0.09	5.00	False	False
'eps': 4, 'min samples': 600	0.47	0.47	0.25	0.07	0.41	3.00	False	False
'eps': 4, 'min samples': 800	0.31	0.31	0.08	0.05	nan	2.00	False	False
'eps': 4, 'min samples': 1000	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'eps': 4.5, 'min samples': 100	0.49	0.49	0.27	0.11	0.11	3.00	False	False
'eps': 4.5, 'min samples': 200	0.45	0.45	0.35	0.12	0.34	4.00	True	True
'eps': 4.5, 'min samples': 300	0.72	0.72	0.54	0.11	-0.29	3.00	True	True
'eps': 4.5, 'min samples': 400	0.70	0.70	0.53	0.10	-0.01	3.00	True	False
'eps': 4.5, 'min samples': 500	0.61	0.61	0.46	0.07	0.47	3.00	True	True
'eps': 4.5, 'min samples': 600	0.50	0.50	0.39	0.08	0.46	3.00	True	True
'eps': 4.5, 'min samples': 800	0.39	0.39	0.17	0.05	nan	2.00	False	False
'eps': 4.5, 'min samples': 1000	0.33	0.33	0.10	0.05	nan	2.00	False	False
'eps': 5, 'min samples': 100	0.49	0.49	0.21	0.10	-0.39	3.00	False	False
'eps': 5, 'min samples': 200	0.44	0.44	0.41	0.09	0.40	4.00	True	True
'eps': 5, 'min samples': 300	0.50	0.50	0.47	0.08	nan	2.00	False	False
'eps': 5, 'min samples': 400	0.71	0.71	0.54	0.11	-0.43	3.00	True	True
'eps': 5, 'min samples': 500	0.61	0.61	0.48	0.08	0.40	3.00	True	True
'eps': 5, 'min samples': 600	0.61	0.61	0.46	0.08	0.45	3.00	True	True
'eps': 5, 'min samples': 800	0.46	0.46	0.31	0.07	0.43	3.00	False	False
'eps': 5, 'min samples': 1000	0.38	0.38	0.17	0.05	nan	2.00	False	False

Table C.115: TEP dataset - OPTICS clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'max eps': 3.2, 'min samples': 50	0.53	0.53	0.11	0.02	0.06	3.00	False	False
'max eps': 3.2, 'min samples': 200	0.42	0.42	-0.00	0.02	0.16	3.00	False	False
'max eps': 3.2, 'min samples': 400	0.33	0.33	-0.04	0.02	nan	2.00	False	False
'max eps': 3.2, 'min samples': 600	0.26	0.26	-0.08	0.02	nan	2.00	False	False
'max eps': 3.2, 'min samples': 800	0.21	0.21	-0.12	0.02	nan	2.00	False	False
'max eps': 3.4, 'min samples': 50	0.41	0.41	0.30	0.02	nan	2.00	False	False
'max eps': 3.4, 'min samples': 200	0.44	0.45	0.07	0.02	0.19	3.00	False	False
'max eps': 3.4, 'min samples': 400	0.37	0.37	0.02	0.02	nan	2.00	False	False
'max eps': 3.4, 'min samples': 600	0.36	0.36	0.00	0.02	nan	2.00	False	False
'max eps': 3.4, 'min samples': 800	0.32	0.32	-0.03	0.02	nan	2.00	False	False
'max eps': 3.6, 'min samples': 50	0.43	0.43	0.22	0.04	0.31	5.00	False	False
'max eps': 3.6, 'min samples': 200	0.54	0.54	0.14	0.02	0.06	3.00	False	False
'max eps': 3.6, 'min samples': 400	0.44	0.44	0.08	0.02	0.21	3.00	False	False
'max eps': 3.6, 'min samples': 600	0.37	0.37	0.06	0.02	nan	2.00	False	False
'max eps': 3.6, 'min samples': 800	0.36	0.36	0.03	0.02	nan	2.00	False	False
'max eps': 3.8, 'min samples': 50	0.44	0.44	0.27	0.05	0.32	5.00	False	False
'max eps': 3.8, 'min samples': 200	0.45	0.45	0.33	0.02	nan	2.00	False	False
'max eps': 3.8, 'min samples': 400	0.50	0.50	0.14	0.02	0.12	3.00	False	False
'max eps': 3.8, 'min samples': 600	0.45	0.45	0.10	0.02	0.17	3.00	False	False
'max eps': 3.8, 'min samples': 800	0.39	0.39	0.10	0.02	nan	2.00	False	False
'max eps': 4, 'min samples': 50	0.45	0.45	0.33	0.06	0.35	5.00	True	True
'max eps': 4, 'min samples': 200	0.49	0.49	0.35	0.02	nan	2.00	True	False
'max eps': 4, 'min samples': 400	0.61	0.61	0.17	0.03	-0.29	3.00	False	False
'max eps': 4, 'min samples': 600	0.53	0.53	0.15	0.02	0.05	3.00	False	False
'max eps': 4, 'min samples': 800	0.48	0.48	0.12	0.02	0.15	3.00	False	False

Table C.116: TEP dataset - OPTICS clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'max eps': 0.6, 'min samples': 50	0.42	0.42	0.27	0.08	-0.02	8.00	True	True
'max eps': 0.6, 'min samples': 200	0.63	0.63	0.53	0.03	0.26	3.00	True	True
'max eps': 0.6, 'min samples': 400	0.54	0.54	0.47	0.03	0.52	3.00	True	True
'max eps': 0.6, 'min samples': 600	0.54	0.54	0.45	0.02	0.52	3.00	True	True
'max eps': 0.6, 'min samples': 800	0.51	0.51	0.37	0.02	nan	2.00	False	False
'max eps': 0.6, 'min samples': 1000	0.51	0.51	0.37	0.02	nan	2.00	False	False
'max eps': 0.6, 'min samples': 1200	0.51	0.51	0.37	0.02	nan	2.00	False	False
'max eps': 0.8, 'min samples': 50	0.41	0.41	0.28	0.07	0.11	9.00	True	True
'max eps': 0.8, 'min samples': 200	0.47	0.47	0.57	0.03	nan	2.00	False	False
'max eps': 0.8, 'min samples': 400	0.61	0.61	0.53	0.03	-0.10	3.00	False	False
'max eps': 0.8, 'min samples': 600	0.54	0.54	0.49	0.03	0.40	3.00	True	True
'max eps': 0.8, 'min samples': 800	0.52	0.52	0.46	0.03	0.43	3.00	False	False
'max eps': 0.8, 'min samples': 1000	0.49	0.49	0.39	0.02	nan	2.00	False	False
'max eps': 0.8, 'min samples': 1200	0.50	0.50	0.38	0.02	nan	2.00	False	False
'max eps': 1, 'min samples': 50	0.42	0.42	0.29	0.07	0.12	8.00	True	True
'max eps': 1, 'min samples': 200	0.46	0.46	0.58	0.03	nan	2.00	True	False
'max eps': 1, 'min samples': 400	0.46	0.46	0.57	0.03	nan	2.00	False	False
'max eps': 1, 'min samples': 600	0.58	0.58	0.53	0.03	0.10	3.00	True	True
'max eps': 1, 'min samples': 800	0.54	0.54	0.51	0.03	0.27	3.00	True	True
'max eps': 1, 'min samples': 1000	0.54	0.54	0.50	0.03	0.18	3.00	False	False
'max eps': 1, 'min samples': 1200	0.48	0.48	0.39	0.02	nan	2.00	False	False

Table C.117: TEP dataset - OPTICS clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’max eps’: 1, ’min samples’: 50	0.42	0.42	0.16	0.15	0.18	6.00	True	True
’max eps’: 1, ’min samples’: 100	0.68	0.68	0.32	0.02	-0.14	3.00	False	False
’max eps’: 1, ’min samples’: 200	0.61	0.61	0.29	0.01	-0.01	3.00	False	False
’max eps’: 1, ’min samples’: 300	0.43	0.43	0.13	0.02	0.34	3.00	True	True
’max eps’: 1, ’min samples’: 400	0.34	0.34	-0.04	0.02	0.26	3.00	False	False
’max eps’: 1, ’min samples’: 500	0.18	0.18	-0.18	0.01	nan	2.00	False	False
’max eps’: 1, ’min samples’: 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’max eps’: 1.2, ’min samples’: 50	0.42	0.42	0.16	0.15	0.18	6.00	False	False
’max eps’: 1.2, ’min samples’: 100	0.49	0.49	0.34	0.03	0.23	3.00	True	True
’max eps’: 1.2, ’min samples’: 200	0.67	0.67	0.32	0.02	-0.24	3.00	False	False
’max eps’: 1.2, ’min samples’: 300	0.60	0.61	0.30	0.01	-0.01	3.00	False	False
’max eps’: 1.2, ’min samples’: 400	0.44	0.44	0.17	0.02	0.36	3.00	False	False
’max eps’: 1.2, ’min samples’: 500	0.41	0.41	0.09	0.02	0.31	3.00	False	False
’max eps’: 1.2, ’min samples’: 600	0.29	0.29	-0.02	0.01	nan	2.00	False	False
’max eps’: 1.5, ’min samples’: 50	0.43	0.43	0.21	0.17	0.17	7.00	True	True
’max eps’: 1.5, ’min samples’: 100	0.46	0.46	0.32	0.04	-0.07	4.00	True	True
’max eps’: 1.5, ’min samples’: 200	0.50	0.50	0.43	0.02	nan	2.00	True	False
’max eps’: 1.5, ’min samples’: 300	0.68	0.68	0.32	0.02	-0.25	3.00	False	False
’max eps’: 1.5, ’min samples’: 400	0.64	0.64	0.31	0.02	-0.24	3.00	False	False
’max eps’: 1.5, ’min samples’: 500	0.53	0.53	0.26	0.02	0.13	3.00	False	False
’max eps’: 1.5, ’min samples’: 600	0.46	0.46	0.21	0.02	0.38	3.00	True	True

Table C.118: TEP dataset - OPTICS clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’max eps’: 1, ’min samples’: 50	0.47	0.47	0.54	0.11	0.58	5.00	False	False
’max eps’: 1, ’min samples’: 100	0.69	0.69	0.53	0.09	-0.20	3.00	False	False
’max eps’: 1, ’min samples’: 200	0.61	0.61	0.47	0.07	0.54	3.00	False	False
’max eps’: 1, ’min samples’: 300	0.59	0.59	0.44	0.07	0.54	3.00	False	False
’max eps’: 1, ’min samples’: 400	0.58	0.58	0.41	0.07	0.54	3.00	False	False
’max eps’: 1, ’min samples’: 500	0.57	0.57	0.40	0.07	0.53	3.00	False	False
’max eps’: 1, ’min samples’: 600	0.55	0.55	0.38	0.07	0.52	3.00	False	False
’max eps’: 1.5, ’min samples’: 50	0.42	0.42	0.54	0.12	0.61	10.00	True	True
’max eps’: 1.5, ’min samples’: 100	0.49	0.49	0.54	0.09	0.62	3.00	True	True
’max eps’: 1.5, ’min samples’: 200	0.71	0.71	0.54	0.09	-0.20	3.00	False	False
’max eps’: 1.5, ’min samples’: 300	0.69	0.69	0.53	0.09	0.22	3.00	False	False
’max eps’: 1.5, ’min samples’: 400	0.60	0.60	0.48	0.07	0.45	3.00	False	False
’max eps’: 1.5, ’min samples’: 500	0.59	0.59	0.46	0.07	0.55	3.00	False	False
’max eps’: 1.5, ’min samples’: 600	0.59	0.59	0.45	0.07	0.55	3.00	False	False
’max eps’: 2, ’min samples’: 50	0.39	0.39	0.44	0.11	0.34	13.00	False	False
’max eps’: 2, ’min samples’: 100	0.47	0.47	0.57	0.11	0.61	4.00	True	True
’max eps’: 2, ’min samples’: 200	0.44	0.44	0.15	0.09	nan	2.00	False	False
’max eps’: 2, ’min samples’: 300	0.43	0.43	0.15	0.09	nan	2.00	False	False
’max eps’: 2, ’min samples’: 400	0.70	0.70	0.54	0.09	-0.28	3.00	False	False
’max eps’: 2, ’min samples’: 500	0.59	0.59	0.49	0.08	0.44	3.00	False	False
’max eps’: 2, ’min samples’: 600	0.58	0.58	0.48	0.07	0.46	3.00	False	False

Table C.119: TEP dataset - OPTICS clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'max eps': 5, 'min samples': 50	0.26	0.26	-0.12	0.13	0.11	16.00	True	True
'max eps': 5, 'min samples': 100	0.65	0.65	0.36	0.09	-0.30	11.00	False	False
'max eps': 5, 'min samples': 150	0.62	0.62	0.32	0.05	-0.18	10.00	False	False
'max eps': 5, 'min samples': 200	0.52	0.52	0.09	0.03	0.20	6.00	False	False
'max eps': 5, 'min samples': 250	0.33	0.33	-0.13	0.04	0.23	5.00	False	False
'max eps': 5, 'min samples': 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 5, 'min samples': 350	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 6, 'min samples': 50	0.26	0.26	-0.12	0.13	0.11	16.00	False	False
'max eps': 6, 'min samples': 100	0.45	0.45	0.19	0.16	-0.01	8.00	True	False
'max eps': 6, 'min samples': 150	0.58	0.58	0.24	0.06	-0.15	7.00	False	False
'max eps': 6, 'min samples': 200	0.68	0.68	0.30	0.05	-0.36	6.00	False	False
'max eps': 6, 'min samples': 250	0.62	0.62	0.22	0.01	-0.11	6.00	False	False
'max eps': 6, 'min samples': 300	0.50	0.50	0.10	0.04	0.19	5.00	False	False
'max eps': 6, 'min samples': 350	0.35	0.35	-0.10	0.05	0.23	5.00	False	False
'max eps': 7, 'min samples': 50	0.26	0.26	-0.12	0.13	0.11	16.00	False	False
'max eps': 7, 'min samples': 100	0.45	0.45	0.19	0.16	-0.01	8.00	False	False
'max eps': 7, 'min samples': 150	0.57	0.57	0.21	0.06	0.25	5.00	False	False
'max eps': 7, 'min samples': 200	0.61	0.61	0.37	0.11	0.18	6.00	True	True
'max eps': 7, 'min samples': 250	0.68	0.68	0.30	0.05	-0.44	6.00	False	False
'max eps': 7, 'min samples': 300	0.66	0.66	0.26	0.03	-0.26	6.00	False	False
'max eps': 7, 'min samples': 350	0.61	0.61	0.21	0.01	-0.10	6.00	False	False
'max eps': 8, 'min samples': 50	0.26	0.26	-0.12	0.13	0.11	16.00	False	False
'max eps': 8, 'min samples': 100	0.45	0.45	0.19	0.16	-0.01	8.00	False	False
'max eps': 8, 'min samples': 150	0.30	0.30	0.08	0.11	0.18	4.00	True	True
'max eps': 8, 'min samples': 200	0.67	0.67	0.38	0.10	0.23	5.00	True	True
'max eps': 8, 'min samples': 250	0.62	0.62	0.33	0.11	0.18	5.00	True	True
'max eps': 8, 'min samples': 300	0.59	0.59	0.32	0.06	0.32	3.00	True	True
'max eps': 8, 'min samples': 350	0.66	0.66	0.27	0.04	-0.32	6.00	False	False

Table C.120: TEP dataset - OPTICS clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
'max eps': 2, 'min samples': 50	0.66	0.66	0.44	0.12	-0.27	7.00	False	False
'max eps': 2, 'min samples': 100	0.61	0.61	0.24	0.10	-0.37	6.00	False	False
'max eps': 2, 'min samples': 200	0.24	0.24	0.02	0.05	nan	2.00	False	False
'max eps': 2, 'min samples': 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 2, 'min samples': 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 2, 'min samples': 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 2, 'min samples': 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 2.5, 'min samples': 50	0.47	0.47	0.24	0.18	0.25	7.00	True	True
'max eps': 2.5, 'min samples': 100	0.68	0.68	0.44	0.09	0.18	5.00	False	False
'max eps': 2.5, 'min samples': 200	0.56	0.56	0.19	0.04	-0.11	5.00	False	False
'max eps': 2.5, 'min samples': 300	0.29	0.29	0.07	0.05	nan	2.00	False	False
'max eps': 2.5, 'min samples': 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 2.5, 'min samples': 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 2.5, 'min samples': 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 3, 'min samples': 50	0.47	0.47	0.31	0.18	0.25	6.00	True	True
'max eps': 3, 'min samples': 100	0.45	0.45	0.35	0.09	0.41	5.00	True	True
'max eps': 3, 'min samples': 200	0.68	0.68	0.51	0.09	0.29	3.00	True	True
'max eps': 3, 'min samples': 300	0.56	0.56	0.18	0.04	-0.12	5.00	False	False
'max eps': 3, 'min samples': 400	0.35	0.35	0.11	0.05	nan	2.00	False	False
'max eps': 3, 'min samples': 500	0.17	0.17	-0.05	0.05	nan	2.00	False	False
'max eps': 3, 'min samples': 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False
'max eps': 4, 'min samples': 50	0.47	0.47	0.31	0.18	0.25	6.00	False	False
'max eps': 4, 'min samples': 100	0.45	0.45	0.39	0.12	0.37	4.00	False	False
'max eps': 4, 'min samples': 200	0.51	0.51	0.47	0.08	nan	2.00	False	False
'max eps': 4, 'min samples': 300	0.68	0.68	0.52	0.10	-0.20	3.00	False	False
'max eps': 4, 'min samples': 400	0.59	0.59	0.43	0.07	0.52	3.00	True	True
'max eps': 4, 'min samples': 500	0.51	0.51	0.22	0.08	-0.05	4.00	False	False
'max eps': 4, 'min samples': 600	0.46	0.46	0.23	0.07	0.41	3.00	False	False
'max eps': 5, 'min samples': 50	0.47	0.47	0.31	0.18	0.25	6.00	False	False
'max eps': 5, 'min samples': 100	0.45	0.45	0.39	0.13	0.38	4.00	True	True
'max eps': 5, 'min samples': 200	0.46	0.46	0.45	0.12	0.36	3.00	True	True
'max eps': 5, 'min samples': 300	0.43	0.43	0.43	0.14	nan	2.00	True	False
'max eps': 5, 'min samples': 400	0.66	0.66	0.52	0.10	-0.18	3.00	True	True
'max eps': 5, 'min samples': 500	0.59	0.59	0.45	0.07	0.49	3.00	True	True
'max eps': 5, 'min samples': 600	0.59	0.59	0.43	0.07	0.49	3.00	True	True

Table C.121: TEP dataset - HDBSCAN clustering - No dimension reduction

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’min cluster size’: 5, ’min samples’: 50	0.43	0.44	0.27	0.06	0.30	7.00	True	True
’min cluster size’: 5, ’min samples’: 100	0.44	0.44	0.36	0.05	0.37	5.00	True	True
’min cluster size’: 5, ’min samples’: 200	0.45	0.45	0.30	0.04	0.37	5.00	True	True
’min cluster size’: 5, ’min samples’: 300	0.48	0.48	0.22	0.04	0.36	4.00	False	False
’min cluster size’: 5, ’min samples’: 400	0.17	0.17	0.47	0.04	0.42	3.00	True	True
’min cluster size’: 5, ’min samples’: 500	0.34	0.34	-0.04	0.02	0.12	3.00	False	False
’min cluster size’: 5, ’min samples’: 600	0.36	0.36	-0.03	0.02	0.12	3.00	False	False
’min cluster size’: 10, ’min samples’: 50	0.43	0.44	0.27	0.06	0.30	7.00	False	False
’min cluster size’: 10, ’min samples’: 100	0.44	0.44	0.36	0.05	0.37	5.00	False	False
’min cluster size’: 10, ’min samples’: 200	0.45	0.45	0.30	0.04	0.37	5.00	False	False
’min cluster size’: 10, ’min samples’: 300	0.48	0.48	0.22	0.04	0.36	4.00	False	False
’min cluster size’: 10, ’min samples’: 400	0.17	0.17	0.47	0.04	0.42	3.00	False	False
’min cluster size’: 10, ’min samples’: 500	0.34	0.34	-0.04	0.02	0.12	3.00	False	False
’min cluster size’: 10, ’min samples’: 600	0.36	0.36	-0.03	0.02	0.12	3.00	False	False
’min cluster size’: 20, ’min samples’: 50	0.44	0.44	0.28	0.05	0.30	6.00	True	True
’min cluster size’: 20, ’min samples’: 100	0.44	0.44	0.36	0.05	0.37	5.00	False	False
’min cluster size’: 20, ’min samples’: 200	0.45	0.45	0.30	0.04	0.37	5.00	False	False
’min cluster size’: 20, ’min samples’: 300	0.48	0.48	0.22	0.04	0.36	4.00	False	False
’min cluster size’: 20, ’min samples’: 400	0.17	0.17	0.47	0.04	0.42	3.00	False	False
’min cluster size’: 20, ’min samples’: 500	0.34	0.34	-0.04	0.02	0.12	3.00	False	False
’min cluster size’: 20, ’min samples’: 600	0.36	0.36	-0.03	0.02	0.12	3.00	False	False
’min cluster size’: 30, ’min samples’: 50	0.44	0.44	0.28	0.05	0.30	6.00	False	False
’min cluster size’: 30, ’min samples’: 100	0.44	0.44	0.36	0.05	0.37	5.00	False	False
’min cluster size’: 30, ’min samples’: 200	0.45	0.45	0.30	0.04	0.37	5.00	False	False
’min cluster size’: 30, ’min samples’: 300	0.15	0.15	0.46	0.04	0.39	3.00	True	True
’min cluster size’: 30, ’min samples’: 400	0.32	0.32	-0.07	0.02	0.09	3.00	False	False
’min cluster size’: 30, ’min samples’: 500	0.34	0.34	-0.04	0.02	0.12	3.00	False	False
’min cluster size’: 30, ’min samples’: 600	0.36	0.36	-0.03	0.02	0.12	3.00	False	False

Table C.122: TEP dataset - HDBSCAN clustering - PCA

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’min cluster size’: 100, ’min samples’: 50	0.56	0.56	0.60	0.07	-0.52	5.00	True	True
’min cluster size’: 100, ’min samples’: 100	0.58	0.58	0.58	0.06	0.06	5.00	False	False
’min cluster size’: 100, ’min samples’: 200	0.58	0.58	0.59	0.04	-0.03	5.00	True	True
’min cluster size’: 100, ’min samples’: 300	0.58	0.58	0.58	0.06	0.27	5.00	True	True
’min cluster size’: 100, ’min samples’: 400	0.58	0.58	0.52	0.02	0.46	4.00	True	True
’min cluster size’: 100, ’min samples’: 500	0.59	0.59	0.51	0.03	0.45	3.00	True	True
’min cluster size’: 100, ’min samples’: 600	0.57	0.57	0.50	0.03	0.48	3.00	True	True
’min cluster size’: 300, ’min samples’: 50	0.56	0.56	0.60	0.07	-0.52	5.00	False	False
’min cluster size’: 300, ’min samples’: 100	0.58	0.58	0.58	0.06	0.06	5.00	False	False
’min cluster size’: 300, ’min samples’: 200	0.58	0.58	0.59	0.04	-0.03	5.00	False	False
’min cluster size’: 300, ’min samples’: 300	0.58	0.58	0.58	0.06	0.27	5.00	False	False
’min cluster size’: 300, ’min samples’: 400	0.60	0.60	0.51	0.03	0.44	3.00	True	True
’min cluster size’: 300, ’min samples’: 500	0.59	0.59	0.51	0.03	0.45	3.00	False	False
’min cluster size’: 300, ’min samples’: 600	0.57	0.57	0.50	0.03	0.48	3.00	False	False
’min cluster size’: 500, ’min samples’: 50	0.59	0.59	0.56	0.05	-0.49	4.00	False	False
’min cluster size’: 500, ’min samples’: 100	0.60	0.60	0.57	0.08	0.02	4.00	False	False
’min cluster size’: 500, ’min samples’: 200	0.64	0.64	0.52	0.03	-0.10	3.00	False	False
’min cluster size’: 500, ’min samples’: 300	0.59	0.59	0.58	0.08	0.25	4.00	True	True
’min cluster size’: 500, ’min samples’: 400	0.60	0.60	0.51	0.03	0.44	3.00	False	False
’min cluster size’: 500, ’min samples’: 500	0.59	0.59	0.51	0.03	0.45	3.00	False	False
’min cluster size’: 500, ’min samples’: 600	0.57	0.57	0.50	0.03	0.48	3.00	False	False
’min cluster size’: 1000, ’min samples’: 50	0.59	0.59	0.56	0.05	-0.49	4.00	False	False
’min cluster size’: 1000, ’min samples’: 100	0.64	0.64	0.51	0.03	-0.03	3.00	False	False
’min cluster size’: 1000, ’min samples’: 200	0.64	0.64	0.52	0.03	-0.10	3.00	False	False
’min cluster size’: 1000, ’min samples’: 300	0.63	0.63	0.52	0.03	0.21	3.00	False	False
’min cluster size’: 1000, ’min samples’: 400	0.60	0.60	0.51	0.03	0.44	3.00	False	False
’min cluster size’: 1000, ’min samples’: 500	0.59	0.59	0.51	0.03	0.45	3.00	False	False
’min cluster size’: 1000, ’min samples’: 600	0.57	0.57	0.50	0.03	0.48	3.00	False	False

Table C.123: TEP dataset - HDBSCAN clustering - MDS

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’min cluster size’: 10, ’min samples’: 50	0.59	0.59	0.38	0.06	-0.26	9.00	True	True
’min cluster size’: 10, ’min samples’: 100	0.56	0.56	0.29	0.04	-0.07	11.00	False	False
’min cluster size’: 10, ’min samples’: 200	0.47	0.47	0.27	0.03	0.14	5.00	False	False
’min cluster size’: 10, ’min samples’: 300	0.18	0.18	0.52	0.04	0.17	3.00	True	True
’min cluster size’: 10, ’min samples’: 400	0.53	0.53	0.25	0.01	0.15	3.00	False	False
’min cluster size’: 10, ’min samples’: 500	0.48	0.48	0.20	0.01	0.23	3.00	True	True
’min cluster size’: 10, ’min samples’: 600	0.45	0.45	0.17	0.01	0.28	3.00	True	True
’min cluster size’: 1000, ’min samples’: 50	0.66	0.66	0.31	0.02	-0.28	3.00	False	False
’min cluster size’: 1000, ’min samples’: 100	0.62	0.62	0.30	0.02	-0.14	3.00	False	False
’min cluster size’: 1000, ’min samples’: 200	0.56	0.56	0.26	0.01	-0.02	3.00	False	False
’min cluster size’: 1000, ’min samples’: 300	0.53	0.53	0.25	0.01	0.15	3.00	False	False
’min cluster size’: 1000, ’min samples’: 400	0.53	0.53	0.25	0.01	0.15	3.00	False	False
’min cluster size’: 1000, ’min samples’: 500	0.48	0.48	0.20	0.01	0.23	3.00	False	False
’min cluster size’: 1000, ’min samples’: 600	0.45	0.45	0.17	0.01	0.28	3.00	False	False
’min cluster size’: 1500, ’min samples’: 50	0.66	0.66	0.31	0.02	-0.28	3.00	False	False
’min cluster size’: 1500, ’min samples’: 100	0.62	0.62	0.30	0.02	-0.14	3.00	False	False
’min cluster size’: 1500, ’min samples’: 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’min cluster size’: 1500, ’min samples’: 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’min cluster size’: 1500, ’min samples’: 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’min cluster size’: 1500, ’min samples’: 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’min cluster size’: 1500, ’min samples’: 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’min cluster size’: 2000, ’min samples’: 50	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’min cluster size’: 2000, ’min samples’: 100	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’min cluster size’: 2000, ’min samples’: 200	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’min cluster size’: 2000, ’min samples’: 300	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’min cluster size’: 2000, ’min samples’: 400	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’min cluster size’: 2000, ’min samples’: 500	0.00	0.00	0.00	0.00	0.00	1.00	False	False
’min cluster size’: 2000, ’min samples’: 600	0.00	0.00	0.00	0.00	0.00	1.00	False	False

Table C.124: TEP dataset - HDBSCAN clustering - ISOMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’min cluster size’: 700, ’min samples’: 50	0.62	0.62	0.57	0.10	0.34	4.00	False	False
’min cluster size’: 700, ’min samples’: 100	0.64	0.64	0.61	0.08	-0.11	5.00	True	True
’min cluster size’: 700, ’min samples’: 200	0.64	0.64	0.58	0.10	0.41	4.00	True	True
’min cluster size’: 700, ’min samples’: 300	0.68	0.68	0.51	0.08	0.41	3.00	True	True
’min cluster size’: 700, ’min samples’: 400	0.67	0.67	0.51	0.08	0.44	3.00	True	True
’min cluster size’: 700, ’min samples’: 500	0.65	0.65	0.50	0.08	0.47	3.00	True	True
’min cluster size’: 700, ’min samples’: 600	0.63	0.63	0.50	0.08	0.33	3.00	False	False
’min cluster size’: 1000, ’min samples’: 50	0.62	0.62	0.57	0.10	0.34	4.00	False	False
’min cluster size’: 1000, ’min samples’: 100	0.64	0.64	0.60	0.11	-0.15	4.00	True	True
’min cluster size’: 1000, ’min samples’: 200	0.64	0.64	0.58	0.10	0.41	4.00	False	False
’min cluster size’: 1000, ’min samples’: 300	0.68	0.68	0.51	0.08	0.41	3.00	False	False
’min cluster size’: 1000, ’min samples’: 400	0.67	0.67	0.51	0.08	0.44	3.00	False	False
’min cluster size’: 1000, ’min samples’: 500	0.65	0.65	0.50	0.08	0.47	3.00	False	False
’min cluster size’: 1000, ’min samples’: 600	0.63	0.63	0.50	0.08	0.33	3.00	False	False
’min cluster size’: 1200, ’min samples’: 50	0.65	0.65	0.50	0.08	0.25	3.00	False	False
’min cluster size’: 1200, ’min samples’: 100	0.67	0.67	0.52	0.08	-0.23	3.00	False	False
’min cluster size’: 1200, ’min samples’: 200	0.67	0.67	0.51	0.08	0.41	3.00	False	False
’min cluster size’: 1200, ’min samples’: 300	0.68	0.68	0.51	0.08	0.41	3.00	False	False
’min cluster size’: 1200, ’min samples’: 400	0.67	0.67	0.51	0.08	0.44	3.00	False	False
’min cluster size’: 1200, ’min samples’: 500	0.65	0.65	0.50	0.08	0.47	3.00	False	False
’min cluster size’: 1200, ’min samples’: 600	0.63	0.63	0.50	0.08	0.33	3.00	False	False

Table C.125: TEP dataset - HDBSCAN clustering - t-SNE

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’min cluster size’: 10, ’min samples’: 50	0.69	0.69	0.27	0.07	-0.16	10.00	False	False
’min cluster size’: 10, ’min samples’: 100	0.53	0.53	0.38	0.09	0.10	2.00	True	True
’min cluster size’: 10, ’min samples’: 200	0.53	0.53	0.38	0.09	0.10	2.00	False	False
’min cluster size’: 10, ’min samples’: 300	0.72	0.72	0.36	0.05	-0.17	4.00	False	False
’min cluster size’: 10, ’min samples’: 400	0.70	0.70	0.42	0.04	-0.16	4.00	True	True
’min cluster size’: 10, ’min samples’: 500	0.67	0.67	0.41	0.04	-0.08	4.00	True	True
’min cluster size’: 10, ’min samples’: 600	0.61	0.61	0.34	0.07	-0.15	5.00	False	False
’min cluster size’: 500, ’min samples’: 50	0.44	0.44	0.39	0.09	-0.52	2.00	True	True
’min cluster size’: 500, ’min samples’: 100	0.53	0.53	0.38	0.09	0.10	2.00	False	False
’min cluster size’: 500, ’min samples’: 200	0.53	0.53	0.38	0.09	0.10	2.00	False	False
’min cluster size’: 500, ’min samples’: 300	0.72	0.72	0.36	0.05	-0.17	4.00	False	False
’min cluster size’: 500, ’min samples’: 400	0.70	0.70	0.42	0.04	-0.16	4.00	False	False
’min cluster size’: 500, ’min samples’: 500	0.67	0.67	0.41	0.04	-0.08	4.00	False	False
’min cluster size’: 500, ’min samples’: 600	0.62	0.62	0.39	0.06	-0.21	4.00	True	True
’min cluster size’: 1000, ’min samples’: 50	0.44	0.44	0.39	0.09	-0.52	2.00	False	False
’min cluster size’: 1000, ’min samples’: 100	0.78	0.78	0.26	0.05	-0.31	4.00	False	False
’min cluster size’: 1000, ’min samples’: 200	0.53	0.53	0.38	0.09	0.10	2.00	False	False
’min cluster size’: 1000, ’min samples’: 300	0.72	0.72	0.36	0.05	-0.17	4.00	False	False
’min cluster size’: 1000, ’min samples’: 400	0.70	0.70	0.42	0.04	-0.16	4.00	False	False
’min cluster size’: 1000, ’min samples’: 500	0.67	0.67	0.41	0.04	-0.08	4.00	False	False
’min cluster size’: 1000, ’min samples’: 600	0.65	0.65	0.41	0.08	-0.22	3.00	True	True

Table C.126: TEP dataset - HDBSCAN clustering - UMAP

Parameters	AMI	V-measure	Silhouette	DB	DBCV	Cluster count	Dominant	Threshold
’min cluster size’: 500, ’min samples’: 50	0.51	0.51	0.12	0.10	-0.27	3.00	False	False
’min cluster size’: 500, ’min samples’: 100	0.64	0.64	0.53	0.03	0.11	5.00	True	True
’min cluster size’: 500, ’min samples’: 200	0.46	0.46	0.48	0.09	-0.28	2.00	True	True
’min cluster size’: 500, ’min samples’: 300	0.70	0.70	0.46	0.06	0.01	4.00	False	False
’min cluster size’: 500, ’min samples’: 400	0.41	0.41	0.44	0.08	0.29	3.00	False	False
’min cluster size’: 500, ’min samples’: 500	0.42	0.42	0.41	0.10	0.25	3.00	True	True
’min cluster size’: 500, ’min samples’: 600	0.61	0.61	0.48	0.08	0.34	3.00	True	True
’min cluster size’: 700, ’min samples’: 50	0.71	0.71	0.36	0.08	0.09	4.00	False	False
’min cluster size’: 700, ’min samples’: 100	0.70	0.70	0.40	0.05	0.03	4.00	False	False
’min cluster size’: 700, ’min samples’: 200	0.46	0.46	0.48	0.09	-0.28	2.00	False	False
’min cluster size’: 700, ’min samples’: 300	0.70	0.70	0.46	0.06	0.01	4.00	False	False
’min cluster size’: 700, ’min samples’: 400	0.41	0.41	0.44	0.08	0.29	3.00	False	False
’min cluster size’: 700, ’min samples’: 500	0.63	0.63	0.48	0.08	0.36	3.00	True	True
’min cluster size’: 700, ’min samples’: 600	0.61	0.61	0.48	0.08	0.34	3.00	False	False
’min cluster size’: 1000, ’min samples’: 50	0.71	0.71	0.36	0.08	0.09	4.00	False	False
’min cluster size’: 1000, ’min samples’: 100	0.70	0.70	0.40	0.05	0.03	4.00	False	False
’min cluster size’: 1000, ’min samples’: 200	0.46	0.46	0.48	0.09	-0.28	2.00	False	False
’min cluster size’: 1000, ’min samples’: 300	0.70	0.70	0.46	0.06	0.01	4.00	False	False
’min cluster size’: 1000, ’min samples’: 400	0.41	0.41	0.44	0.08	0.29	3.00	False	False
’min cluster size’: 1000, ’min samples’: 500	0.63	0.63	0.48	0.08	0.36	3.00	False	False
’min cluster size’: 1000, ’min samples’: 600	0.61	0.61	0.48	0.08	0.34	3.00	False	False
’min cluster size’: 1200, ’min samples’: 50	0.51	0.51	0.12	0.10	-0.27	3.00	False	False
’min cluster size’: 1200, ’min samples’: 100	0.70	0.70	0.40	0.05	0.03	4.00	False	False
’min cluster size’: 1200, ’min samples’: 200	0.46	0.46	0.48	0.09	-0.28	2.00	False	False
’min cluster size’: 1200, ’min samples’: 300	0.70	0.70	0.46	0.06	0.01	4.00	False	False
’min cluster size’: 1200, ’min samples’: 400	0.65	0.65	0.48	0.08	0.33	3.00	True	True
’min cluster size’: 1200, ’min samples’: 500	0.63	0.63	0.48	0.08	0.36	3.00	False	False
’min cluster size’: 1200, ’min samples’: 600	0.61	0.61	0.48	0.08	0.34	3.00	False	False