Erasmus University Rotterdam

Erasmus School of Economics

Bachelor Thesis Econometrie en Operationele Research

Extending the ”Comparative Analysis of Clustering Quality” Framework by Renjith et al. (2021) using Joint Dimensionality Reduction and Clustering

Romi Versluis (533557)

Abstract

With the amount of generated data growing at an unprecedented rate, gaining insight into this data is becoming increasingly valuable. This study uses two datasets of differing nature to investigate the framework by Renjith et al. (2021) on performing dimensionality reduction before clustering high-dimensionality data to improve clustering quality. This study finds that joint dimensionality reduction and clustering methods LDA-Km and CDR would be a valuable addition to this framework.

Supervisor:	drs. J Durieux
Second assessor:	dr. M Khismatullina
Date final version:	2nd july 2023

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics or Erasmus University Rotterdam.

1 Introduction

In today's digital age, the amount of data generated is growing at an unprecedented rate. From social media interactions to online transactions and sensor readings, data is being produced in vast quantities every second. Clustering and dimensionality reduction together provide a powerful technique that is useful for handling Big Data because it allows us to deal with a rising challenge: the curse of dimensionality.

The curse of dimensionality (Bellman, 1961) refers to the problem of having high dimensionality in the data, which can lead to sparsity and computational inefficiency. In other words, as the number of features increases, the amount of data needed to accurately represent the data grows exponentially, which makes it difficult to analyse and cluster.

Dimensionality reduction techniques, such as investigated in this study, can be used to address this problem by reducing the number of features while preserving the most relevant information. By reducing the dimensionality of the data, we can make it easier to analyze and visualize, while also reducing the computational complexity of clustering algorithms.

Furthermore, using dimensionality reduction can help to improve the accuracy and interpretability of clustering results. By reducing the dimensionality of the data and clustering in the reduced feature space, we can identify clusters that are more compact and separable, which can lead to better clustering results. Additionally, the reduced feature space can be easier to interpret, making it easier to identify the underlying patterns and relationships in the data. In summary, we can achieve faster, more efficient, and more accurate clustering results.

This study builds upon a research by Renjith et al. (2021) which provides a framework for dimensionality reduction and clustering for handling high-dimensionality data. Renjith et al. (2021) find that for the data set under consideration, t-SNE outperforms the other investigated methods on both k-means and agglomerative hierarchical clustering (AGNES).

To further complement this framework, this study investigates adaptive dimension reduction using linear discriminant analysis and k-means clustering (LDA-Km) (Ding & Li, 2007) and the clustering and dimensionality method (CDR) with both Reduced k-means (RKM) and Factorial k-means (FKM) (van de Velden et al., 2019). The methods proposed in the framework by Renjith et al. (2021) rely on separately applying dimensionality reduction before clustering. These sequential methods can improve the computational efficiency, but the results deducted from the dimension reduction process may not be optimal for the clustering process, so that some researchers believe that the separation of dimension reduction and clustering may result in worse clustering performance. Intuitively, if clustering is embedded into the process of dimension reduction, the performance of clustering may be improved. Methods like these try to find the optimal structure of data in the low-dimensional feature space for clustering. (Long et al., 2021).

Therefore, this study investigates whether LDA-Km and/or CDR would be a good addition to the framework by Renjith et al. (2021) by comparing them not only to PCA + k-means, but also to the other methods considered in this framework as t-SNE was found to be the best-performing method of the current framework under the experimental circumstances. The research question for this study is: "How well do joint dimensionality reduction and clustering methods LDA-Km and CDR perform in dimensionality reduction and clustering compared to the methods considered in the framework by Renjith et al. (2021) using the same data set?" and consequently "Would LDA-Km and/or CDR be a valuable addition to the framework by Renjith et al. (2021)?" which is evaluated by also considering data of a different nature.

2 Literature

As mentioned before, this study builds upon a research by Renjith et al. (2021) which provides a framework for dimensionality reduction and clustering for handling high-dimensionality data. K-means clustering and AGNES are used as candidate clustering methods and the dimensionality reduction techniques evaluated include principal component analysis (PCA), independent component analysis (ICA), t-distributed stochastic neighbor embedding (t-SNE) and locally linear embedding (LLE). Renjith et al. (2021) find that for the data set under consideration, t-SNE outperforms the other investigated methods with both k-means and AGNES.

There are several methods that perform joint clustering and dimensionality reduction, two of the more-frequently used methods are LDA-Km (Ding & Li, 2007) and CDR (van de Velden et al., 2019; Vichi et al., 2019) . CDR is based on a model which uses factorial k-means and reduced k-means and simultaneously uses PCA as dimensionality reduction technique (Vichi et al., 2019; van de Velden et al., 2019). Both models use k-means as clustering algorithm, yet they differ in dimensionality reduction technique. Between PCA and linear discriminant analysis (LDA), it is generally believed that algorithms based on LDA are superior to those based on PCA (Borade & Adgaonkar, 2011). PCA and LDA are both essentially linear techniques where objectives can be solved by solving an eigenvalue problem with the corresponding eigenvector defining the hyperplane of interest (Xanthopoulos et al., 2013). However, PCA can be used to find a subspace whose basis vectors best represent the original data, while LDA searches for vectors that best discriminate among classes (rather than those that best describe the data). Even though regular LDA is used for classification problems and uses target labels, LDA-Km does not require labels, just as PCA, yet has the same objective. Intuitively, the objective of LDA seamlessly fits with clustering, together with the belief that LDA algorithms are superior to PCA (Martinez & Kak, 2001), LDA-Km could possibly provide new insights to the framework by Renjith et al. (2021). Both LDA-Km and CDR are investigated and compared in this study to measure both their possible contribution to the framework by Renjith et al. (2021).

Ding & Li (2007) combines LDA and k-means clustering into a framework to ultimately select the most discriminative subspace. The research uses k-means clustering to generate class labels and uses LDA to select a subspace. The clustering process is thus integrated with the subspace selection process and the data are simultaneously clustered and the subspaces selected. The effectiveness of this method is proven by looking at variants of the method and their correspondence with earlier approaches and is found to be effective.

Vichi et al. (2019) use a combination of PCA and one of various forms of k-means clustering simultaneously in order to determine the best possible clustering. The study distinguishes between k-means, Reduced k-means and Factorial k-means as possible clustering methods, which might prove useful in finding the highest possible clustering quality. The advantages of the method is shown by applying it in a simulation study and two empirical example analyses.

However LDA-Km and CDR seem promising in outperforming both k-means and PCA+k-means (Ding & Li, 2007) , they are not yet compared to t-SNE. In Renjith et al. (2021), PCA is found to be outperformed by t-SNE for both k-means clustering and AGNES, therefore comparing these methods to t-SNE would provide a better insight in the performance power of LDA-Km and CDR as it provides a higher benchmark for these specific data sets.

3 Data

Initially, as part of the aim of this study is to replicate the study by Renjith et al. (2021), the same data set is used. This data set is also used for the application of LDA-Km as this provides possibilities to compare LDA-Km to the methods used in the known framework.

The data set used is an open-source set called the "Jester data set". This set contains of 4.1 million observations where each of the observations is an anonymous rating of a joke on a scale from -10 to 10 which are gathered from April 1999 to May 2003. In total there are 100 jokes which are rated by 73,421 readers.

The data pre-processing and in particular the missing data handling is not mentioned in the paper. What is mentioned, is that a sample of 5000 is used for computing the clusters, but the selection process of this sample is not stated. Chances are that the 5000 points selected do not contain missing values and do not need further processing, however this is not certain. For this study, the data containing missing values is removed and out of the remaining observations, 5000 points are selected at random, using a set seed.

Furthermore, next to the ratings to the 100 jokes, the data set contains an additional column which states the amount of jokes that that specific user has rated. Following the procedure executed by Renjith et al. (2021), this column is removed.

Renjith et al. (2021) state in their conclusion that the nature of the data set has a strong influence on the performance of the clustering and dimensionality reduction algorithms. Therefore after completing the replication and evaluation of the performance of LDA-Km and CDR on this data, the same process is be repeated on a data set with a different nature, to further elaborate on the strength of the framework.

The data set chosen to further show the strength of the framework, is a data set concerning air line customer satisfaction. This data set consists of a mix of categorical and continuous data, rather than only continuous data. This dataset is a collection of 103904 responses to a survey regarding the satisfaction of that specific customer. The customer is asked to provide some personal information, for example their gender and their flying class, and is asked to rate a selection of aspects regarding their experience with the airline. The ratings range from 0 to 5 where 1 indicates the worst rating, 5 the highest and 0 is filled in whenever the service is not applicable to said interviewee. Alongside these observations, several flight-related information is granted. Specifics on the data set can be found in Appendix A.

The data preprocessing has been conducted in the same manner as used for the Jester data set. First, the observations containing missing values are removed, after which 5000 random data points are selected. As not all dimensionality reduction methods support non-numerical data, the categorical attributes are mapped into numbers and as t-SNE requires distinct data points, non-distinct observations are removed. More thorough elaboration on the data preprocessing is provided in Appendix B.

In order to provide a more complete insight into the effectiveness of the used models, two data sets having a different nature are used. The Jester data set used in Renjith et al. (2021) contains solely continuous data, whereas the Airline Customer Satisfaction data set also includes categorical data. The Clustrd package differentiates between categorical data and continuous data and is therefore able to handle non-numeric data. Furthermore, LLE, t-SNE and CDR can deal with non-numeric data, yet do not have a specified method for categorical data. However, both PCA and ICA are unable to handle non-numeric data, therefore the categorical data is mapped into numbers. A full insight into the chosen mapping can be found in Appendix B.

4 Methodology

In this section, all methods that are used in conducting this study are discussed. Starting off with four methods that are solely used for dimensionality reduction in Section 4.1, followed by k-means clustering and agglomerative hierarchical clustering (AGNES) in Section 4.2 (Renjith et al., 2021). Furthermore the methods for performing joint dimensionality reduction and clustering are mentioned in Section 4.3. In order to measure the performance of all methods stated, internal cluster validation criteria are selected which are illustrated in Section 4.4.

4.1 Dimensionality Reduction

In order to handle increasing quantity of data and to improve clustering results, several dimensionality reduction techniques are used as a data processing stage before clustering the data. This study compares PCA, ICA, LLE and t-SNE with two joint dimensionality reduction techniques (CDR and LDA-Km).

4.1.1 Principal Component Analysis

PCA is a feature projection approach for dimensionality reduction, and it extracts a new set of attributes called principal components as linear combination of original attributes. In PCA, the principal components are populated with the first new attribute explains the maximum variation. The second new attribute attempts to explain the remaining variation and so on. Typically, it is observed that more than 60% of the variation in a data set is explained by initial four principal components in PCA. The algorithm for determining the principal components is stated below.

Algorithm:	Principal Component Analysis (Levada, 2021)
Step 1:	Compute the sample mean and the sample covariance matrix by $\mu_x = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\Sigma_x = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \mu_x)(x_i - \mu_x)^T$
Step 2:	Compute the eigenvalues and eigenvectors of Σ_x
Step 3:	Define the transformation matrix $T = [w_1, w_2, ..., w_d$ with the d eigenvectors associated to the d largest eigenvalues.
Step 4:	Project the data X into the PCA subspace: $y_i = Tx_i$ for $i = 1, 2, ...n$
Step 5:	Return Y

In order to determine the optimal amount of components to include in the dimensionality reduction, the explained variance is used. Once the PCA is conducted, the variance per principal component is known and the percentage of total variance of each component can be computed as follows:

$$EV_i = \frac{var(PC_i)}{\sum_i (var(PC_i))} \times 100 \tag{1}$$

where $var(PC_i)$ denotes the variance of principal component i.

4.1.2 Independent Component Analysis

ICA provides a way to find a linear coordinate system such that the resulting signals are as statistically independent from each other as possible. In contrast to correlation-based transformations such as PCA, ICA not only decorrelates the signals (2nd-order statistics) but also reduces higher-order statistical dependencies. The full algorithm on the computation of the used ICA method is stated below.

Algorithm:	Independent Component Analysis (fastICA) (Shlens, 2014)
Step 1:	Subtract off the mean of the data in each dimension.
Step 2:	Whiten the data by decorrelating using the eigenvectors of the covariance of the data (E) and normalize the data using the eigenvalues (D). $x_w = (D^{-\frac{1}{2}}E^T)x$ where x_w is the whitened version of observed data x.
Step 3:	Identify final rotation matrix that optimizes statistical independence by solving $V = \arg\min_V \sum_i H[(Vx_w)_i]$ where V is a rotation matrix and $H[(Vx_w)_i]$ is the entropy of $(Vx_w)_i$.

As the ICA Algorithm includes the PCA steps, the optimal amount of components of ICA is the same as determined for PCA. Therefore for the determination of the number of components for ICA, Equation 1 is used.

4.1.3 Locally Linear Embedding

Locally Linear Embedding (LLE) is a nonlinear dimensionality reduction technique that aims to find a lower-dimensional representation of the data while maintaining the local pairwise relationships between neighboring data points. If we denote the original dataset as $X = [x_1, x_2, ..., x_N] \in$

$R^{D\times N}$ with $x_i \in R^D$ the i'th sample for $i = 1, 2, ..., N$ where N is the size of the sample. LLE maps X into $Y = [y_1, y_2, ..., y_N] \in R^{d\times N}$, where d is far smaller than D. Mathematically, LLE involves the following steps:

Algorithm:	Locally Linear Embedding (Roweis & Saul, 2000)
Step 1	Find the neighbors for each point. For a point x_i, $X^i = [x_i^1, x_i^2, ..., x_i^k] \in R^{D\times k}$ are the neighbors of x_i where k is the number of neighbors. Pairwise similarity in the data is calculated via Euclidean distance and the k most similar points to x_i are chosen as neighbors. $d(x_i, \mu_j) = \|x_i - \mu_j\|_2$.
Step 2:	Reconstruct with linear weights. The weights $W_i = [w_{i1}, w_{i2}, ..., w_{ik}]' \in R^{k\times 1}$ are calculated using the following least squares problem. $\min \epsilon(W) = \sum_i \lvert \vec{X}_i - \sum_j W_{ij}\vec{X}_j \rvert^2$
Step 3:	Map each observation $\vec{X}_i$ into lower-dimensional vector $\vec{Y}_i$ by choosing coordinates $\vec{Y}_i$ to minimize the cost function $\Phi(Y) = \sum_i \lvert \vec{Y}_i - \sum_j W_{ij}\vec{Y}_j \rvert^2$.

The value of k in k-nearest-neighbors is determined according to the method proposed by Kayo (2006), where the error term $\epsilon(W)$ is computed for multiple values of k. The value of k which minimizes the error term, is chosen as optimal value of k.

4.1.4 t-Distributed Stochastic Neighbor Embedding

t-SNE is a non-linear dimensionality reduction technique which uses neighbors to find similarity among the observations. Data points are considered neighbors if their distance is less than a certain threshold value. It is a stochastic method which uses the t-distribution as probability distribution function, which tends to a normal distribution when using a large sample size. The conditional probability is given by:

$$p_{j|i} = \frac{\exp(-\|x_i - x_j\|^2/2\sigma_i^2}{\sum_{k\neq i}\exp(-\|x_i - x_k\|^2/2\sigma_i^2} \tag{2}$$

where σ_i^2 denotes the squared standard deviation, which is also known as the variance. Mathematically, t-SNE involves the following steps:

where the joint probabilities used in the algorithm are computed as follows:

$$q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-}1}{\sum_{k\neq l}(1 + \|y_i - y_l\|^2)^{-}1} \tag{3}$$

and the formula for determining the gradient is:

$$\frac{\delta C}{\delta y_i} = 4\sum_j (p_{ij} - q_{ij}(y_i - y_j)(1 + \|y_i - y_j\|^2)^{-}1 \tag{4}$$

```
Algorithm:  t-Distributed Stochastic Neighbor Embedding
            (van der Maaten & Hinton, 2008)
Data:       data set $X = x_1, x_2, ..., x_n$
            cost function parameter perplexity, Perp
            optimization parameters: number of iterations T, learning rate $\eta$
            , momentum $\alpha(t)$
Result      low-dimensional data representation $Y^{(0)} = y_1, y_2, ..., y_n$.
Step 1:     compute pairwise affinities $p_{j|i}$ with perplexity Perp (using Equation 2)
Step 2:     set $p_{ij} = \frac{pj|i+pi|j}{2n}$
Step 3:     sample initial solution $Y^{(0)} = y_1, y_2, ..., y_n$ from $N(0, 10^{-4}I)$
            for t=1 to T do
Step 4:     compute low-dimensional affinities $q_{ij}$ (using Equation 3)
Step 5:     compute gradient $\frac{\delta C}{\delta Y}$ (using Equation 4)
Step 6:     set $Y^{(t)} = Y^{(t-1)} + \eta\frac{\delta C}{\delta Y} + \alpha(t)(Y^{(t-1)} - Y^{(t-2)})$
```

In the framework by Renjith et al. (2021), t-SNE is chosen as a dimensionality reduction technique before clustering, therefore it is repeated in this study. Originally, t-SNE is a visualization tool that forms a low-dimensional space to show a 2D representation. The t-SNE method is not designed to make an embedding based on the clusters in the real data. Therefore, one may find t-SNE to produce high-quality clusters by chance, but might also encounter that existing clusters in the data are disregarded. In this study, the perplexity parameter is tuned such that the values of the internal cluster validation indices were indicating good results, yet this method does require iterative tuning and is therefore not easily applicable to other cases.

4.2 Clustering

In this research, popular clustering techniques k-means and Agglomerative Nesting (AGNES) are used to form clusters based on the original data and the transformed data produced by the dimensionality reduction techniques. Having determined the clusters for both the original data and the transformed data, enables to verify the effectiveness of data dimensionality. This is verified using multiple criteria, which are further enlightened in Section 4.4.

In order to obtain the optimal amount of clusters to use in this research, the R package NbClust was consulted (Charrad et al., 2014). This package uses 26 different indices which all determine the optimal amount of clusters given the data set, based on their individual conditions. After this, the number of clusters which is returned by most of the indices is chosen to further use in this study.

4.2.1 k-means Clustering

k-means is an unsupervised clustering method that aims to partition data into k clusters such data points in the same cluster are close and data points in the different clusters are as far away as possible. Formally, the k-means algorithm aims to minimize the within-cluster sum of squares (WCSS), which is defined as the sum of squared distance between each data point and is denoted as

$$\underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = \underset{\mathbf{S}}{\arg\min} \sum_{i=1}^{k} |S_i| \operatorname{Var} S_i \tag{5}$$

where S denotes the set of points in cluster i and μ denotes the centroid of the corresponding cluster.

```
Algorithm: k-means Clustering (Hartigan & Wong, 1979)

Step 1:  Randomly initialize K centroids, denoted as $\mu_1, \mu_2, ..., \mu_k$.
Step 2:  For each data point $x_i$, calculate the Euclidean distance to each centroid:
         $d(x_i, \mu_j) = \|x_i - \mu_j\|_2$.
Step 3:  Assign each data point to the nearest centroid, minimizing the distance:
         $c_i = arg\min_j(d(x_i, \mu_j))$.
Step 4:  Recalculate the centroids as the mean of the data points assigned to each cluster:
         $\mu_j = (1/|C_j|)\sum_{i=1}^{T}(x_i)$ if $i \in C_j$, where $|C_j|$ represents the number of data points
         assigned to cluster j.
Step 5:  Repeat Steps 2,3 and 4 until convergence, when the centroids do not change
         significantly .
```

Given the nature of the algorithm, k-means clustering can converge to local optima, so running the algorithm multiple times with different initializations is often recommended to improve the results.

4.2.2 Agglomerative Hierarchical Clustering

Agglomerative nesting (AGNES) is a hierarchical clustering algorithm that aims to group similar objects or data points together based on their pairwise similarity or dissimilarity. The algorithm uses a bottom up approach, which means that it starts with each data point as an individual cluster and iteratively merges clusters until a desired stopping condition is met. This results in a nested hierarchy of clusters that can be represented as a dendrogram. This full dendogram is later cut such that it encapsulates the determined optimal amount of clusters. A pseudo algorithm for AGNES is provided below.

```
Algorithm: Agglomerative Nesting (Ramos Emmendorfer & de Paula Canuto, 2021)

Step 1:  for a set $X = [x_1, x_2, ..., x_n$ with n objects. Create set of partitions
         $\Pi = [\pi_0, \pi_1, ..., \pi_{n-1}$ where each observation has its own cluster.
Step 2:  Merge 2 clusters P and Q for which the linkage L(P, Q) is lowest among all
         pairs in $\pi_{k-1}$.
Step 3:  Replace P,Q with R, leaving $\pi_{k-1} \leftarrow \pi_k$.
Step 4:  Repeat until the number of clusters in $\pi_k$ is 1.
```

4.3 Joint Dimensionality Reduction and Clustering

In this subsection, the dimensionality reduction and clustering methods that are tested to extend the framework by Renjith et al. (2021) are introduced. Starting with the CDR method in Section 4.3.1, which is split into Reduced k-means and Factorial k-means, followed by LDA-Km in Section 4.3.2.

4.3.1 Clustering and Dimensionality Reduction

CDR is a method that proposes to combine the PCA loss function and the FKM loss function. In this combination, one can tweak the value of α to determine the importance of each of the two loss functions. The proposed loss function by Vichi et al. (2019) is as follows:

$$f(F, A, U, Y) = \alpha\|X - FA'\|^2 + (1-\alpha)\|XAA' - UYA'\|^2 \tag{6}$$

where X is the data matrix, F denotes an $I \times Q$ matrix of component scores, and it is used that $\|XAA^T - UYA^T\|^2 = \|XA - UY\|^2$. A is a $J \times Q$ loading matrix for Q dimensions with the property that $A^T A = I$. U indicates an $I \times K$ binary matrix indicating the memberships of the objects to K clusters and Y is the $K \times Q$ cluster centroids matrix.

Altering the value of α provides different loss functions, which has the specific characteristics:

- Setting $\alpha = 0$ gives the FKM loss function.
- Setting $\alpha = 0.5$ gives the RKM loss function.
- Setting $\alpha = 1$ gives the regular PCA loss function.

In this study, both FKM and RKM are investigated and compared.

Based on the type of data used in the experiment, CDR offers differentiated methods for categorical data and continuous data. Out of all investigated methods, CDR is the only method that differentiates between different natures of data (adhering to the framework by Renjith et al. (2021)). To provide a more complete research on the effectiveness of the investigated models, in this study an additional data set is considered containing categorical data. As a result, both CDR methods (”Cluspca” and ”Clusmca”) (van de Velden et al., 2019) are incorporated in this study.

4.3.2 Linear Discriminant Analysis and k-means Clustering

Regular LDA is a method used for conducting supervised classification, meaning that each observation requires a label. This label serves as a value for a specific target and is used as additional information for the dimensionality reduction next to the other features. The combination of LDA and k-means is an unsupervised method and therefore labels are no requirement and are not used in the dimensionality reduction and clustering.

LDA and k-means both optimize the same objective function. They both minimize the within-class scatter matrix and maximize the between-class scatter matrix. In LDA, we use the total scatter, between-class scatter and within-class scatter matrices, respectively S_t, S_b and S_w:

$$S_t = \sum_{i=1}^{n} x_i x_i^T \tag{7}$$

$$S_b = \sum_{k} n_k m_k m_k^T \tag{8}$$

$$S_w = \sum_k \sum_{i \in C_k} (x_i - m_k)(x_i - m_k)^T \tag{9}$$

where $S_t = S_w + S_b$. The objective function of k-means clustering is

$$J_K = trS_w = tr(S_t - S_b) \tag{10}$$

and trS_w is used to indicate the trace of S_w.

The main goal of this framework is to optimize the LDA objective function, which has the same aim as the k-means objective function. The LDA objective function is

$$\max_{U,H} tr \frac{U^T S_b U}{U^T S_w U} \tag{11}$$

where U denotes the projectional matrix, also named the LDA directions. The matrix $H = \{0,1\}^{n \times K}$ is the cluster indicator: $H_{ik} = 1$ if x_i belongs to the k-th cluster, and $H_{ik} = 0$ otherwise. The U matrix consist of the LDA directions.

Algorithm:	Adaptive LDA-guided k-means Clustering (Ding & Li, 2007)
Step 1:	Set K = number of clusters, Set d = K − 1 the dimension of the subspace.
Step 2:	Compute PCA on X to obtain initial U.
Step 3:	Do step LDA-Km(1) to obtain H, see Equation 12.
Step 4:	Do step LDA-Km(2) to obtain U.
Step 5:	Go to step 3 until convergence.

where LDA-Km(1) involves solving for H while fixing U in:

$$\max_{h} \frac{trU^T S_b U}{trU^T S_w U} \tag{12}$$

which is exact same as the LDA objective function stated in Equation 11.

LDA-Km(2) refers to solving for U while fixing H. U is given by the standard LDA procedure.

What the algorithms essentially does, is initialize U and then alternately perform k-means and LDA in the subspace until convergence has been reached.

4.4 Internal Cluster Validation

In order to evaluate the performance of the dimensionality reduction and clustering, several indices are selected. Each of the indices has their own criteria and therefore the four indices together provide a representative overview of the performance. The aim and the formula for each of the indices are stated below.

4.4.1 Silhouette Index

The silhouette value is a measure of how similar an object is to its own cluster compared to other clusters. The calculation of the index is stated in Equations 13, 14 and 15. The silhouette ranges from − 1 to + 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate. If many points have a low or negative value, then the clustering configuration may have too many or too few clusters.

$$a(i) = \frac{1}{|C_I| - 1} \sum_{j \in C_I, i \neq j} d(i,j) \tag{13}$$

$$b(i) = \min_{J \neq I} \frac{1}{|C_J|} \sum_{j \in C_J} d(i,j) \tag{14}$$

$$I = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{15}$$

4.4.2 Dunn Index

The Dunn index is another measure of clustering quality. For this index, higher values indicate that the within-cluster variance is low and the between-cluster variance is high. Therefore higher values mean better clustering. The formula in stated below in Equation 16

$$I = \min_{1 \leq i \leq c}\{ \min_{1 \leq j \leq c, j \neq i}\{\frac{\delta(C_i, C_j)}{\max_{1 \leq k \leq c}\{\Delta(C_k)\}}\}\} \tag{16}$$

where c is the number of defined clusters and $\delta(C_i, C_j)$ denotes the intercluster distance between cluster i and j.

4.4.3 Calinski–Harabasz Index

Same as the other indices, the Calinski-Harabasz measure of clustering validation. Here the inter-cluster quality is estimated based on the distances from the data points in a cluster to its cluster centroid and between-cluster quality is determined by the distance of the cluster centroids from the global centroid. Higher values of this index stand for higher clustering quality. To calculate the value, Equation 17 is used

$$I = \left[\frac{\sum_{k=1}^{K} n_k \left\|c_k - c\right\|^2}{K-1}\right] / \left[\frac{\sum_{k=1}^{K} \sum_{i=1}^{n_k} \left\|x_i - c_k\right\|^2}{N-K}\right] \tag{17}$$

for which n_k and c_k respectively indicate the number of data points and the centroid of the k'th cluster. In the formula, c is the notation used for indicating the global centroid. K denotes the number of clusters on dataset $X = [x_1, x_2, ..., x_N$, given this formulation N denotes the number of data points.

4.4.4 Davies–Bouldin Index

As all indices, the Davies-Bouldin index is a metric for evaluating the formed clusters. Opposed to the three other indices, lower Davies-Bouldin values indicate more appropriate clustering. The formula is as follows:

$$I = \frac{1}{k} \sum i = 1^k \max_{i \neq j}\{\frac{\Delta(X_i) + \Delta(X_j)}{\delta(X_i, X_j)}\} \tag{18}$$

here, $\Delta(X_i)$ denotes the within cluster scatter of cluster i. $\delta(X_i, X_j)$ accounts for the separation between the i'th and the j'th cluster.

5 Results

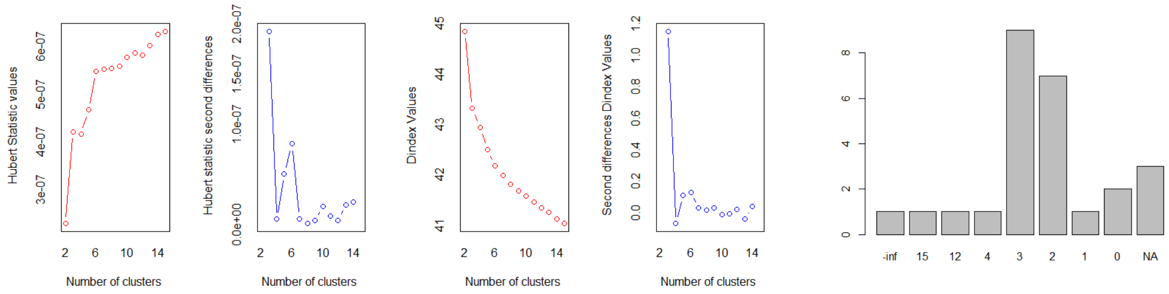
This Section provides an overview of all findings in this research. Starting off with the determination of the optimal amount of clusters in Section 5.1, followed by insights regarding the formed clusters with and without dimensionality reduction in Section 5.2. Concluded by an evaluation of the clusters in Section 5.3. In Section 5.4, the same process is repeated yet includes categorical data.

5.1 Optimal amount of clusters

In order to abide by the experimental analysis of Renjith et al. (2021) as much as possible, a specific linkage method (Ward squared) was selected to ensure that the optimal amount of clusters would be determined as being three. The results of each of the evaluated indices is indicated in Table 1 along with further detailed results shown in Figure 1.

Table 1: Determination of optimal count of clusters

Indices	Scott	Dindex	McClain	Marriot	Ball	KL	SDbw	Hartigan	TrCovW	TraceW	Friedman	CCC	Ratkowsky	CH	Beale	Cindex	Rubin	DB	Frey	Hubert	PtBiserial	Dunn	SDindex	Silhouette	Duda	PseudoT2
Jester	-inf	15	12	4	3	3	3	3	3	3	3	3	3	2	2	2	2	2	2	2	1	0	0	NA	NA	NA



(a)

(b)

(c)

Figure 1: Detailed insights into the optimal amount of cluster determination

5.2 Formed clusters

Having determined the optimal amount of clusters being three, the clustering is first executed on both the data that is not reduced in dimensionality. The two dimensional view of this clustering is shown on the left in both Figure 2 and Figure 3 in order to make a clear comparison to the dimensionality reduction with both linear and non-linear methods. Figure 2 depicts the clustering results of both k-means and AGNES using linear methods PCA and ICA. Figure 3 provides an overview of the clusters formed using non-linear methods LLE and t-SNE.

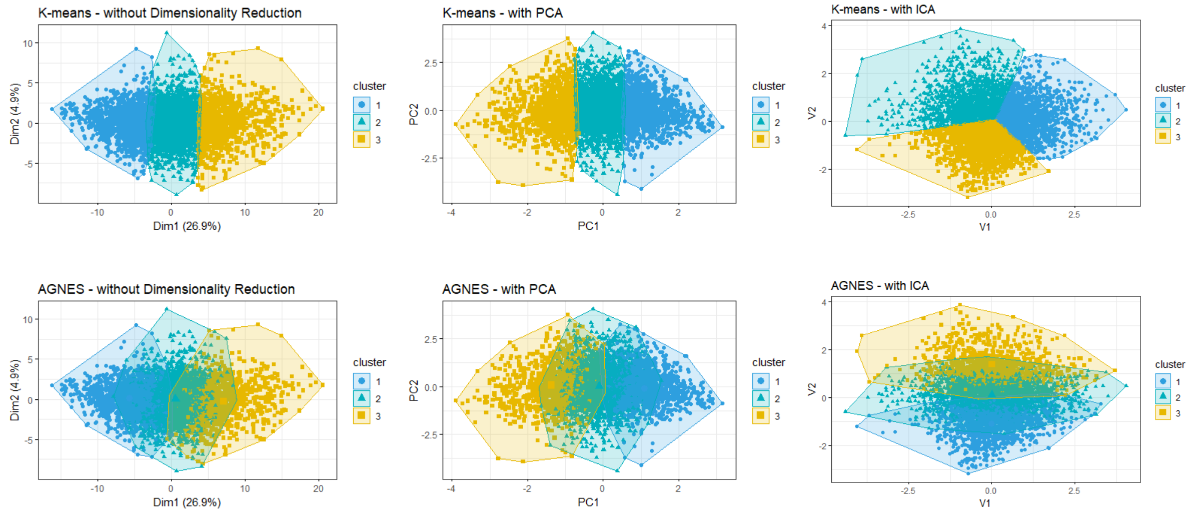


Figure 2: Two-dimensional view of clusters formed with linear dimensionality reduction techniques alongside those formed without dimensionality reduction

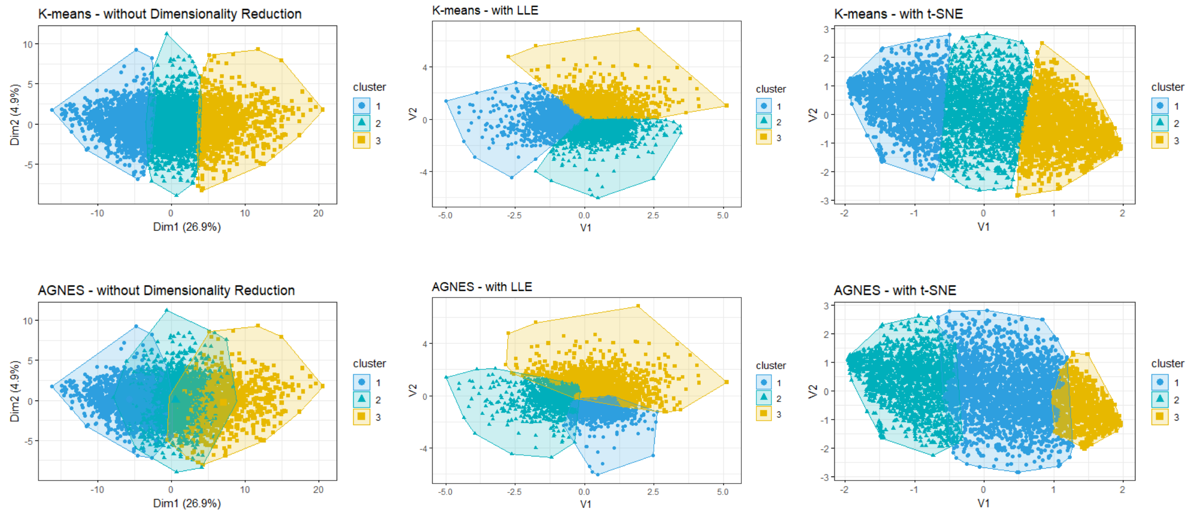


Figure 3: Two-dimensional view of clusters formed with non-linear dimensionality reduction techniques alongside those formed without dimensionality reduction

To gain insight into the quality of the clusters shown above, the clusters are evaluated using four internal evaluation criteria. These are elaborated on in Section 5.3.

5.3 Cluster Evaluation

The performance of the formed clusters after dimensionality reduction is measured by means of four internal evaluation criteria. The indices used are the Silhouette index (Section 4.4.1), Dunn index (Section 4.4.2), Calinski-Harabasz index (Section 4.4.3) and the Davies-Bouldin index (Section 4.4.4. Figure 4 shows the impact of the different dimensionality reduction techniques for both k-means and AGNES clustering.

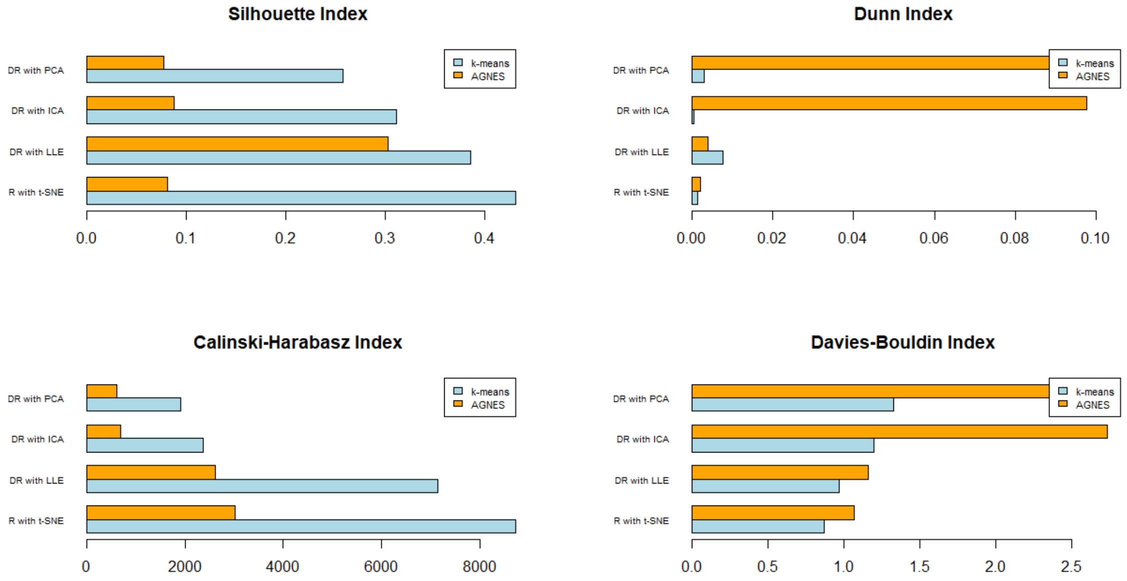


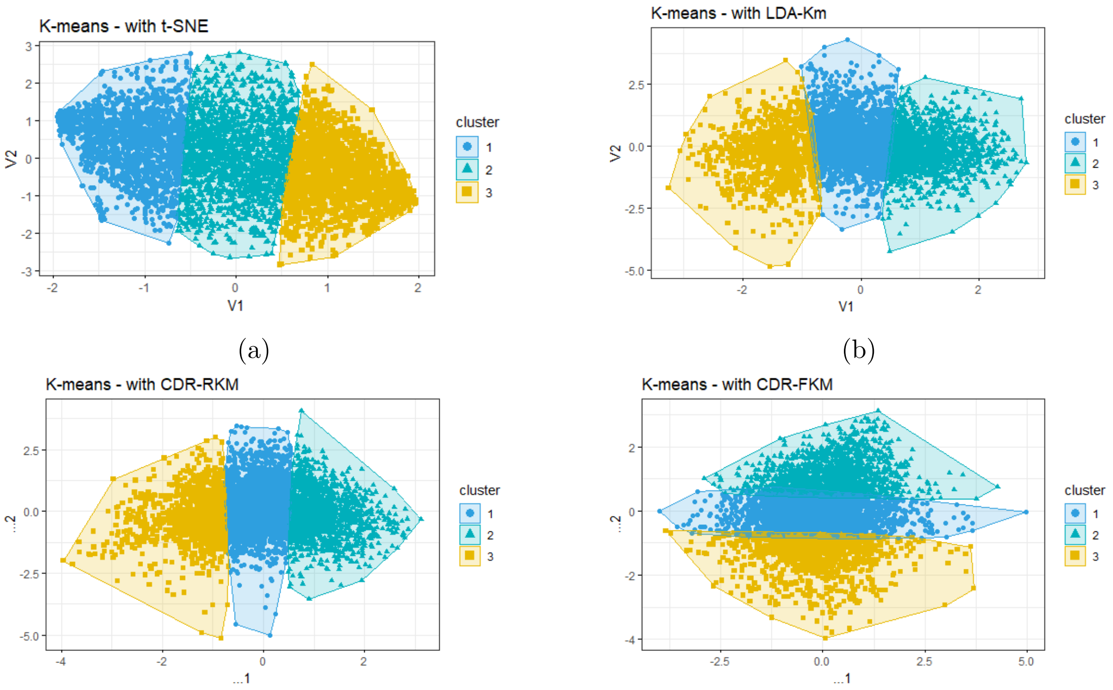
Figure 4: Comparison of PCA, ICA, LLE and t-SNE by means of internal evaluation criteria

For the Silhouette index, a higher value indicates a higher quality of clustering. In this study, as depicted in Figure 4, t-SNE has the best value for k-means and is therefore the best performing technique in this case, followed by LLE. As for AGNES, LLE outperforms all other methods.

Higher values of the Dunn index indicate better clustering. Therefore, PCA and ICA score significantly high for agnes, which substantially deviates from the findings by k-means, where PCA and ICA score relatively low and LLE scores highest. For the Calinski-Harabasz index, higher values stand for better clustering quality. For both k-means and AGNES, t-SNE outperforms the other methods. The same counts for Davies-Bouldin, where lower numbers indicate better clustering, t-SNE outperforms all three other methods for both k-means and AGNES.

When investigating the findings above, it is found that t-SNE is superior to the other models in most of the cases for this study. To further investigate the contribution of joint dimensionality reduction and clustering methods (Section 4.3), t-SNE is compared to LDA-Km, CDR-RKM and CDR-FKM.

Below in Figure 5, the four plots show a two-dimensional overview of the formed clusters. First plot, Figure 5a, shows dimensionality reduction with t-SNE, the method that is found to be best performing in this experiment, having investigated the models used in (Renjith et al., 2021). To possibly further extend and enhance this framework, several joint dimensionality reduction and clustering methods are considered. In the second plot 5b, one can see the clustering results formed by the LDA-Km method. Plot three 5c gives a two-dimensional view of clustering with CDR-RKM. Lastly, Figure 5d shows the clustering formed by the CDR-FKM method.



(a) (b)

(c) (d)

Figure 5: Two-dimensional view of clusters formed with t-SNE, LDA-Km, CDR-RKM and CDR-FKM

Having computed the dimensionality reduction and clustering, the different methods are evaluated using internal cluster validation indices, as explained in Section 4.4. The values of these indices are summarized in the bar plots below in Figure 6.

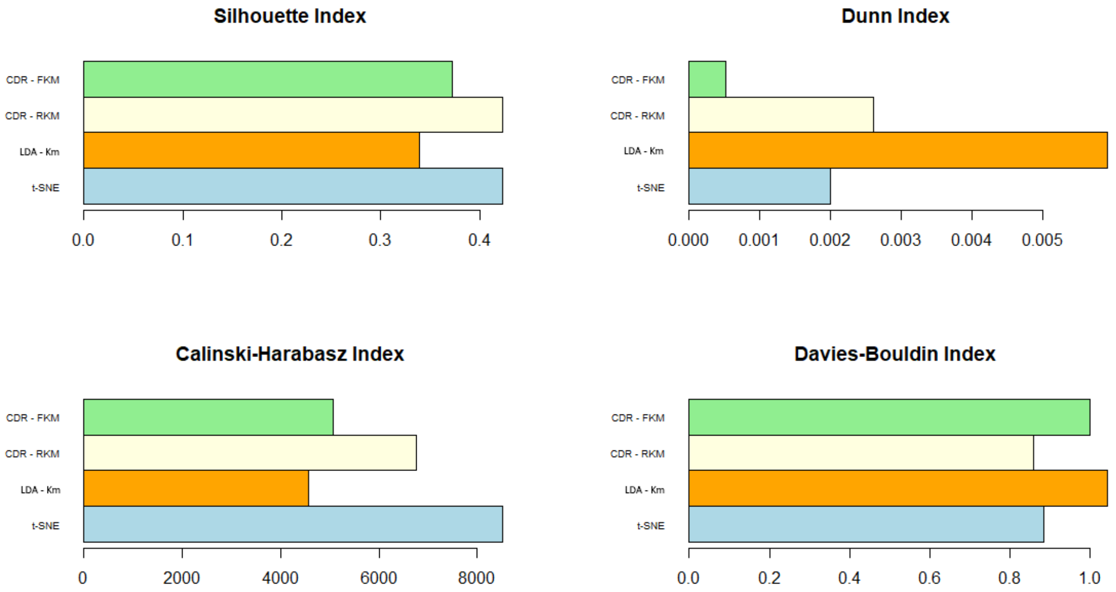


Figure 6: Comparison of t-SNE to LDA-Km and CDR by means of internal evaluation criteria

The Silhouette index is highest for the best clustering quality. Both CDR-RKM and t-SNE score higher than the other two methods, yet do not differ much. For the Dunn index, higher values mean better formed clusters and it is visible in Figure 6 that LDA-Km outperforms all other methods. The Calinski-Harabasz index has a higher score whenever the performance of the clustering is better. One can see that t-SNE scores best for this index, followed by CDR-RKM. For Davies-Bouldin index, lower values indicate better clustering. Here LDA-Km is the best performing dimensionality reduction and clustering method, closely followed by CDR-FKM. One can conclude from explanation above and as depicted in Figure 6, best performing dimensionality reduction methods vary for the different indices and there is not one uniform best method given the results of the indices.

5.4 Considering categorical and continuous data

Since Renjith et al. (2021) conclude that the nature of the data highly impacts the effectiveness of the methods, the process performed on the Jester data set is repeated using a data set which has a different nature, namely the Airline Customer Satisfaction data which contains categorical data.

Figure 7 provides a two-dimensional overview of the clusters formed by both k-means and AGNES without dimensionality reduction and with applying the methods PCA and ICA before clustering. The same is done for LLE and t-SNE, which is shown in Figure 8. As can be seen in this Figure, the optimal amount of clusters determined is 5. This is computed by NbClust which uses the Ward Squared method to calculate 26 different indices, of which 5 determined 5 to be the optimal amount of clusters. The Ward Squared method is chosen following the study by Renjith et al. (2021) as explained in Section 5.1.

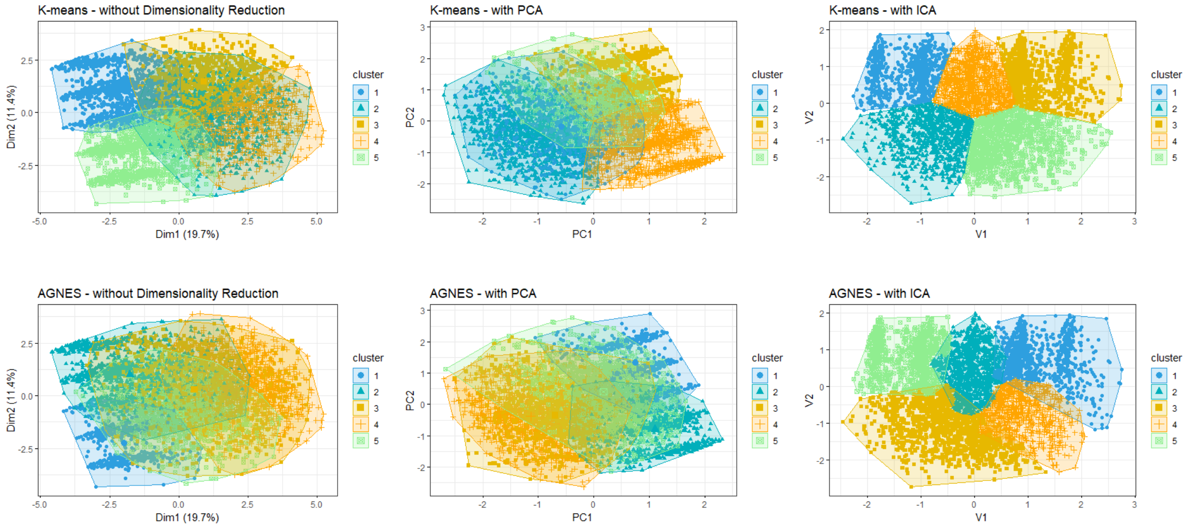


Figure 7: Two-dimensional view of clusters formed with linear dimensionality reduction techniques alongside those formed without dimensionality reduction

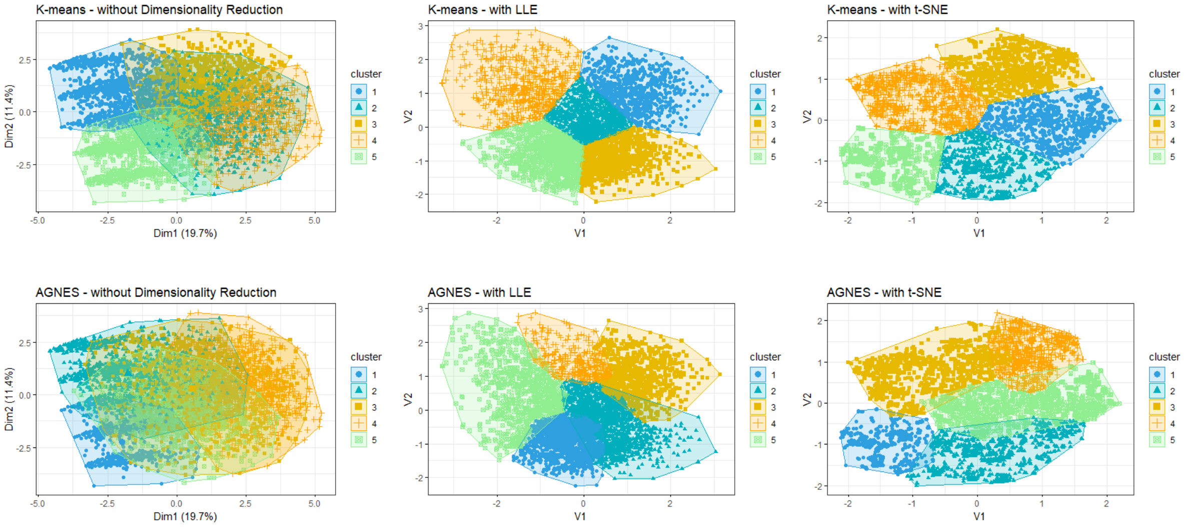


Figure 8: Two-dimensional view of clusters formed with non-linear dimensionality reduction techniques alongside those formed without dimensionality reduction

As can be seen in Figures 7 and 8, PCA is not designed to handle categorical data, this method does not perform optimally. There are alternatives for performing PCA on categorical data, one of which is CATPCA by Linting & van der Kooij (2012), which is specifically designed for dealing with non-numeric data. As the framework by Renjith et al. (2021) is considered as reference throughout this paper, the choice is made to compare the models included in this framework, and to therefore use regular PCA instead of any possibly better alternative.
Having performed the dimensionality reduction and clustering, the formed clusters are evaluated using internal cluster validation indices, as explained in Section 4.4. The values for all dimensionality reduction methods for both k-means and AGNES are shown in the plots below.

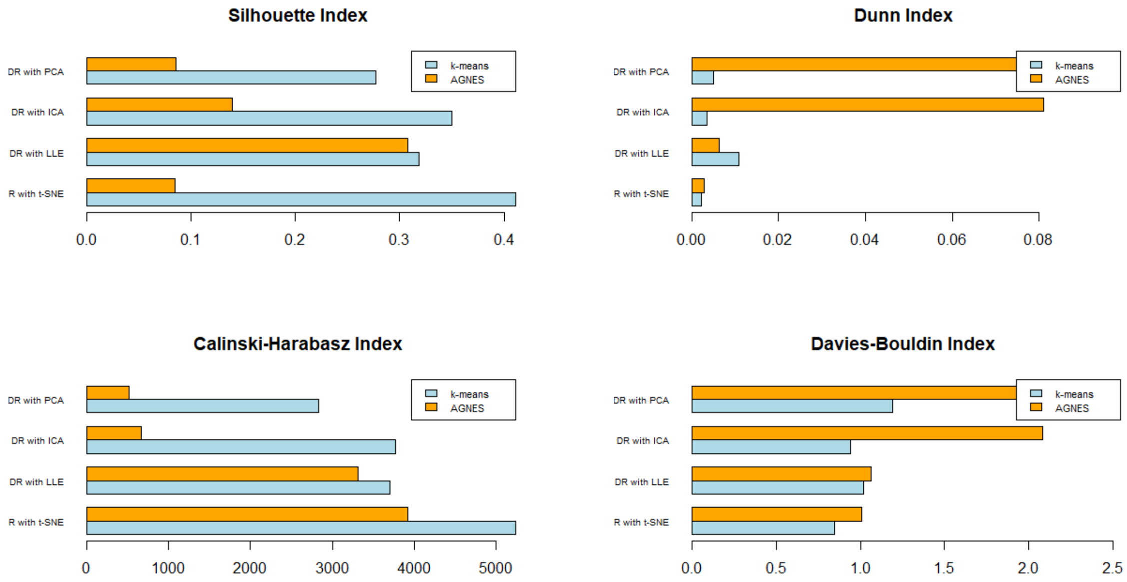
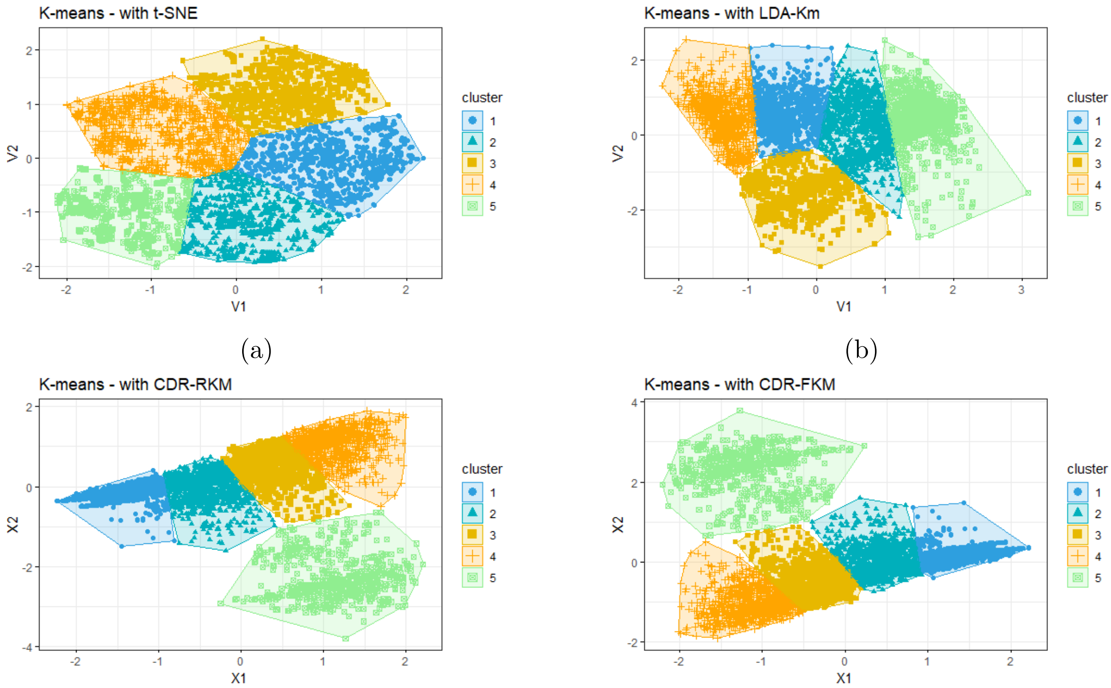


Figure 9: Comparison of PCA, ICA, LLE and t-SNE by means of internal evaluation criteria

As fully elaborated on in Section 4.4, for the Silhouette index, Dunn index and Calinski-Harabasz index, higher values indicate better quality of clustering. The Davies-Bouldin index value is lower whenever the clusters are better formed. As can be deducted from Figure 9, for the Silhouette index, t-SNE outperformes the other methods when considering k-means. For AGNES, LLE outperforms PCA, ICA and t-SNE. The values of the Dunn index show that both PCA and ICA are the best performing methods. The Calinski-Harabasz index values indicate that t-SNE is the method that appears to have the best clustering quality. For both k-means and AGNES, t-SNE has the lowest Davies-Bouldin index value and is therefore the best performing model according to this index.

Although not completely convincing, t-SNE is indicated as being the best dimensionality reduction method by most of the criteria compared to the other methods. Therefore t-SNE will be the method that is compared to the joint dimensionality reduction and clustering methods investigated to possibly enhance the framework by Renjith et al. (2021).

The comparison between t-SNE and joint dimensionality reduction and clustering methods LDA-Km, CDR-RKM and CDR-FKM are showcased in Figure 10. In this Figure, the two-dimensional view of the formed clusters is shown for all four of the models.



(a)

(b)

(c)

(d)

Figure 10: Two-dimensional view of clusters formed with t-SNE, LDA-Km, CDR-RKM and CDR-FKM

What stands out in 8, is that the two-dimensional plots of the CDR-RKM and the CDR-FKM are almost identical, except flipped. This indicates that the PCA part of Equation 6 is not much represented and the k-means part is of most importance. As a result, the data is robust to most variations in the value of α. After having computed the dimensionality reduction and clustering, the results are evaluated using the internal validation indices discussed in Section 4.4.

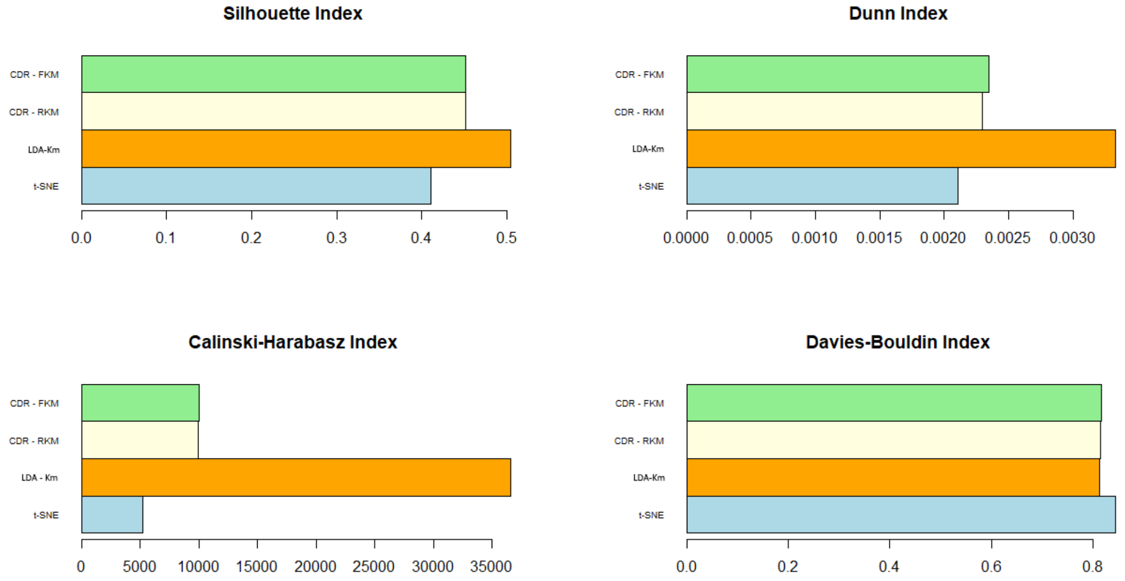


Figure 11: Comparison of t-SNE, LDA-Km, CDR-RKM and CDR-FKM by means of internal evaluation criteria

As can be seen in Figure 11, the values for all indices seem to be relatively close. Yet for the Silhouette index, LDA-Km performs best as this method produces the highest value. Both CDR-RKM and CDR-FKM outperform t-SNE. Looking at the Dunn index, LDA-Km outperforms the other methods, having the highest value. For the Calinksi-Harabasz index, higher values indicate higher clustering quality, therefore LDA is best performing according to this index. Clustering methods with lower Davies-Bouldin values, have higher quality. Values for CDR-FKM, CDR-RKM and LDA-Km do not significantly differ, however appear to be better than the clustering formed using t-SNE.

One can argue that LDA-Km seems to be the best fit for this specific experiment, with CDR-FKM and CDR-RKM being good alternatives according to most indices. What stands out is that t-SNE is outperformed according to all the methods, showing that joint dimensionality reduction and clustering is a valuable addition to this experiment.

6 Conclusion and Discussion

The primary goal of this study was to investigate the framework proposed by Renjith et al. (2021) and research whether joint dimensionality reduction and clustering methods LDA-Km and CDR would be a valuable addition to this framework.

The study uses the Jester data set as well as the Airline Customer Satisfaction data set to firstly measure the clustering quality of the methods proposed in the framework by Renjith et al. (2021). The resulting clusters are evaluated using four internal cluster validation indices, showing that t-SNE outperforms the other methods in both of the data sets, even though this is highly dependent on the value of the perplexity parameter. Afterwards, the process is repeated now comparing t-SNE to the joint dimensionality reduction and clustering methods LDA-Km, CDR-RKM and CDR-FKM.

What is found in this study, is that there is no uniform best method among these four, however it can be concluded that t-SNE found to be outperformed by the three joint dimensionality reduction and clustering methods for both of the data sets by most of the computed indices.

Renjith et al. (2021) stated in their conclusion that the nature of the data has an impact on the quality of the different investigated methods, which this research is completely in line with. However, from the finding above, one can conclude that the joint dimensionality reduction and clustering methods are proven be a valuable addition to the framework by Renjith et al. (2021), as for the experiments conducted in this study, the investigated extension methods are proven to outperform the proposed methods.

For further research, it would be valuable to research the performance of joint dimensionality reduction and clustering methods which incorporate a non-linear dimensionality reduction method. Since t-SNE is a non-linear method which outperformed all sequential methods and linear methods PCA and LDA are used in LDA-Km and CDR, a combination of joint dimensionality reduction and clustering with a non-linear method could be performing even better. A possible method that can be considered is an extension of LDA-Km to a Nonlinear Case, which is proposed by Ding & Li (2007). This proposition uses kernels to implement the non-linear transformation as a linear one by mapping to a higher dimensional space.

7 Appendix

References

Bellman, R. (1961). *Adaptive control processes: A guided tour. (A RAND Corporation Research Study).* Princeton, N. J.: Princeton University Press, XVI, 255 p. (1961).

Borade, S. N., & Adgaonkar, R. P. (2011). Comparative analysis of pca and lda. In *2011 international conference on business, engineering and industrial applications* (p. 203-206).

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2014). Nbclust: An r package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, *61*(6), 1–36. doi: 10.18637/jss.v061.i06

Ding, C., & Li, T. (2007). Adaptive dimension reduction using discriminant analysis and k-means clustering. In (p. 521–528). New York, NY, USA: Association for Computing Machinery.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108.

Kayo, O. (2006). Locally linear embedding algorithm - extensions and applications. Universitatis Ouluensis, Oulu, Finland.

Levada, A. L. M. (2021). Pca-kl: a parametric dimensionality reduction approach for unsupervised metric learning. *Advances in Data Analysis and Classification*, *15*(4), 829–868. doi: 10.1007/s11634-020-00434-3

Linting, M., & van der Kooij, A. (2012). Nonlinear principal components analysis with catpca: A tutorial. *Journal of Personality Assessment*, *94*(1), 12-25. doi: 10.1080/00223891.2011.627965

Long, Z.-Z., Xu, G., Du, J., Zhu, H., Yan, T., & Yu, Y.-F. (2021). Flexible subspace clustering: A joint feature selection and k-means clustering framework. *Big Data Research*, *23*, 100170.

Martinez, A., & Kak, A. (2001). Pca versus lda. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(2), 228-233. doi: 10.1109/34.908974

Ramos Emmendorfer, L., & de Paula Canuto, A. M. (2021). A generalized average linkage criterion for hierarchical agglomerative clustering. *Applied Soft Computing*, *100*, 106990. doi: https://doi.org/10.1016/j.asoc.2020.106990

Renjith, S., Sreekumar, A., & Jathavedan, M. (2021). A comparative analysis of clustering quality based on internal validation indices for dimensionally reduced social media data. *Advances in Artificial Intelligence and Data Engineering*, 1047–1065.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323-2326. doi: 10.1126/science.290.5500.2323

Shlens, J. (2014). A tutorial on independent component analysis. *arXiv preprint arXiv:1404.2986*.

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

van de Velden, M., Iodice, D. A., & Yamamoto, M. (2019). Special feature: dimension reduction and cluster analysis. *Behaviormetrika*, *46*, 239-241.

Vichi, M., Vicari, D., & Kiers, H. A. L. (2019). Clustering and dimension reduction for mixed variables. *Behaviormetrika*, *46*, 243-269.

Xanthopoulos, P., Pardalos, P. M., & Trafalis, T. B. (2013). Linear discriminant analysis. In *Robust data mining* (pp. 27–33). New York, NY: Springer New York.

A Characteristics of the Airline Customer Satisfaction data set

The Airline Customer Satisfaction data set consists of several types of input, being:

- Personal questions, such as Age and Gender.

- Travel-specific questions, such as Type of Travel and Flight Distance (known before the flight).
- Service-related questions, such as Cleanliness and Ease of Online Booking (evaluated after the flight).

The data set contains several attributes which are known before flight. These are, 'Age', 'Gender', 'Customer Type' (Loyal / Non-loyal), 'Flight Distance', 'Type of Travel' (for personal or business reasons) and 'Class' (Eco, Eco Plus or Business). The sample characteristics are represented in Table 2.

Table 2: Sample Characteristics

Characteristics	Statistics
Age	Mean (39.38)
Gender	Male (49%), Female (51%)
Customer Type	Loyal (82%), Disloyal (18%)
Flight Distance	Mean (1189.45 miles)
Type of Travel	Personal (31%), Business (69%)
Class	Eco (45%), Eco Plus (7%), Business (48%)

Above the known pre-flight variables, the data set contains attributes regarding service quality extracted from a conducted survey. For each of the service related questions, the interviewed customers are asked to rate a specific service on a scale from one to five, one being the lowest score and five the highest. If a specific service was not applicable for the customer, a zero was filled in. The percentages per score and mean of the answers on each topic are provided in Table 3 below.

Table 3: Distribution of Service Attributes (in %) and its Mean

Rating	0	1	2	3	4	5	Mean
Inflight wifi service	2.99	17.17	24.86	24.90	19.05	11.04	2.73
Departure/Arrival time convenient	5.10	14.92	16.55	17.29	24.59	21.56	3.06
Ease of Online booking	4.32	16.87	23.12	23.53	18.84	13.33	2.76
Gate location	0.00	16.90	18.73	27.50	23.51	13.36	2.98
Food and drink	0.10	12.35	21.16	21.46	23.44	21.47	3.20
Online boarding	2.34	10.29	16.85	20.98	29.61	19.93	3.25
Seat comfort	0.00	11.62	14.34	17.99	30.57	25.48	3.44
Inflight entertainment	0.01	12.01	16.97	18.42	28.32	24.27	3.36
On-board service	0.00	11.43	14.13	21.98	29.71	22.76	3.38
Leg room service	0.45	9.96	18.79	19.34	27.71	23.74	3.35
Baggage handling	0.00	6.97	11.09	19.86	35.98	26.11	3.63
Checkin service	0.00	12.41	12.41	27.38	27.96	19.84	3.30
Inflight service	2.99	17.17	24.86	24.90	19.05	11.04	3.64
Cleanliness	0.01	12.82	15.53	23.65	26.16	21.84	3.29

B Preprocessing of the Airline Customer Satisfaction data set

- The target label "Satisfaction Level" is removed from the data.
- The variables ID and number of interviewee are deleted from the data.
- observations including missing values are removed.
- identical observations are removed such that all preserved observations are different.
- Gender: Male = 1 and Female = 2
- Customer Type: Disloyal = 1 and Loyal = 2
- Age: $\text{Age} \leq 18 = 1$, $18 < \text{Age} \leq 24 = 2$, $24 < \text{Age} \leq 35 = 3$, $35 < \text{Age} \leq 65 = 4$, $\text{Age} > 65 = 5$
- Type of Travel: personal = 1, business = 2
- Class: Eco = 1, Eco Plus = 2, Business = 3
- Flight Distance (miles): $\text{Distance} \leq 900 = 1$, $900 < \text{Distance} \leq 1500 = 2$, $\text{Distance} > 1500 = 3$. This categorization is based on the classification from the United Airlines MileagePlus Program.
- For all categorical variables regarding service, up the value with one to change the range from 0-5 to 1-6.
- Departure Delay (in min): $\text{delay} \leq 15 = 1$, $15 < \text{delay} \leq 180 = 2$, $180 < \text{delay} \leq 240 = 3$, $\text{delay} > 240 = 4$. This categorization corresponds with EU regulations on compensation for delayed flights. EU regulation was chosen as there is no US standard for these compensations.
- Arrival Delay (in min): $\text{delay} \leq 15 = 1$, $15 < \text{delay} \leq 180 = 2$, $180 < \text{delay} \leq 240 = 3$, $\text{delay} > 240 = 4$. Similar categorization to the departure delay.