Unsupervised Learning Project

Juliana Restrepo Tobar
Universidad EAFIT
Medellin, Colombia
jrestrepot@eafit.edu.co

Olga Lucía Quintero Montoya
Universidad EAFIT
Medellin, Colombia
oquinte1@eafit.edu.co

***Abstract*—In this study, we undertake a detailed analysis of the Iris dataset using various unsupervised learning techniques. Our approach includes an exploratory data analysis (EDA) employing pair plots, biplots with Principal Component Analysis (PCA), and Uniform Manifold Approximation and Projection (UMAP). Following the EDA, we extensively test multiple clustering algorithms, each with different hyperparameter configurations, and evaluate their performance using internal indices. This report not only explores the Iris dataset through advanced data analysis methods but also provides a comparative evaluation of diverse unsupervised learning algorithms, highlighting their effectiveness in revealing inherent data patterns.**

I. Introduction

Clustering algorithms, a cornerstone of unsupervised learning, play a pivotal role in discerning hidden patterns and structures within unlabeled data. This paper delves into the realm of clustering, emphasizing the evaluation and comparative analysis of various algorithms in this field. Our journey begins with a nod to the origins of unsupervised learning, tracing its evolution from the early concepts of hierarchical clustering in the 1950s to the advanced methods of today.

While hierarchical clustering laid the groundwork by introducing the concept of organizing data into nested structures, the evolution of computational capabilities has propelled the field into new dimensions. The emergence of algorithms like K-means in the 1960s and the introduction of techniques such as Principal Component Analysis (PCA) marked significant milestones, enhancing the ability to segment data based on similarities and dimensional characteristics.

The advent of big data and machine learning in recent decades has further underscored the importance of unsupervised learning. Clustering algorithms have diversified, finding applications across various sectors—from image segmentation in tech to customer segmentation in marketing. This era has witnessed the rise of sophisticated techniques like Fuzzy c-Means, DBSCAN, and spectral clustering, alongside dimensionality reduction methods like UMAP, adept at handling complex, high-dimensional data.

In this project, we focus on a systematic evaluation of these clustering algorithms, scrutinizing their performance across different data representations: original, compressed (via autoencoders and UMAP), and expanded (using autoencoder techniques). Our methodology involves a detailed exploration of algorithms such as connected components, naive distance clustering, fuzzy c-Means, and K-Means, with a special emphasis on leveraging mountain and subtractive clustering for initializing centers. By conducting a comprehensive assessment through a blend of internal and external indices, our goal is to unravel the nuances of each algorithm, revealing their strengths, limitations, and optimal application scenarios. This endeavor is more than an academic exercise; it's a quest to harness the true potential of unsupervised learning in extracting meaningful insights from data.

II. Methodology

A. Data

The Iris dataset is a well-known and frequently used dataset in the field of machine learning and statistics. It was introduced by the British biologist and statistician Ronald A. Fisher in 1936. This dataset consists of 150 samples (n) of iris flowers, representing three different species: setosa, versicolor, and virginica. For each sample, four features (m) are measured: sepal length, sepal width, petal length, and petal width, all in centimeters. The primary purpose of the Iris dataset is to serve as a benchmark for classification algorithms, making it a fundamental resource for tasks like pattern recognition and machine learning model training and evaluation.

B. Descriptive data analysis

1) Basic statistics: In this step, basic statistical measures such as mean, median, standard deviation, and quartiles will be computed for each of the four features (sepal length, sepal width, petal length, and petal width) in the Iris dataset. These statistics will provide a summary of the central tendency and variability of the data.

2) Histograms: Histograms will be generated for each feature to visualize the distribution of data. Each histogram will represent the frequency or count of data points within specified bins or intervals for that particular feature. This will give insights into the data's distribution and any potential skewness. It also serves to assess if there are different populations within the data.

3) Biplot: We will create a biplot by performing Principal Component Analysis (PCA) to reduce the dataset's dimensionality from the original four features (sepal length, sepal width, petal length, and petal width) to just two principal components (PC1 and PC2). This plot allows us to visually assess the contributions of each original feature to the variation in the data and how data points cluster or separate based on these features.

4) Pair plot: For further analysis we'll use a pair plot, sometimes referred to as a scatterplot matrix. This plot will display scatterplots for all possible pairs of features, allowing for the examination of pairwise relationships and correlations within the dataset. This can help identify any strong correlations or patterns between features. This plot differs from the others because we'll use a separate color for each species, which are our target labels.

5) UMAP: UMAP, which stands for Uniform Manifold Approximation and Projection will be applied to the Iris dataset. UMAP is a dimensionality reduction technique used to visualize high-dimensional data in a lower-dimensional space while preserving the essential structure and relationships among data points. The UMAP projection will be created, and the resulting lower-dimensional representation of the Iris dataset will be used for further analysis or visualization.

C. Clustering Algorithms

For a dataset X with n observations, we have the following algorithms.

1) Mountain Clustering: Mountain clustering uses a density measure called the mountain function to estimate cluster centers in a dataset. It divides the space into clusters resembling mountains, where minimizing the metric within cluster members and their centroids while maximizing the distance between centroids is crucial. This method involves using Gaussian-like kernels and Euclidean distance. Initially, the search space is gridded, and finer grids yield better cluster estimations but increase computational costs. The mountain function's height represents the number of data points around a node. Finding cluster centers is an iterative process, eliminating the effect of current centers to identify new ones. The algorithm mimics absorbing nearby clusters based on density, aiding in understanding spatial notions, metrics, and optimization in learning from data. Mountain clustering serves as a preliminary step for more complex algorithms needing initial cluster centers. Parameters σ and β influence the mountain function's construction and update: σ influences height and smoothness, while β affects the appearance of the new function.

The algorithm involves creating an equally spaced grid on the entire data space. We set V as the set of all of the points where the grid lines intersect each other. Then, we compute the value of the *initial* mountain function m_1 at each point $v \in V$ with equation 1:

$$m_1(v) = \sum_{j=1}^{n} e^{\left(\frac{-||v-x_j||^2}{2\sigma^2}\right)} \tag{1}$$

After having computed the equation 1 for every v, we designate the cluster center c_1 as the point v where the function m_1 reaches the maximum value. Finally, we start an iterative process of computing the value of the mountain function m_{i+1} at each point $v \in V$ with equation 2

$$m_{i+1}(v) = m_i(v) - m_i(c_i)e^{\left(\frac{-||v-c_i||^2}{2\beta^2}\right)} \tag{2}$$

We do this iteratively process until the algorithm finds a center it already found before or until it reaches the maximum number of iterations (100).

In our case, we will create a grid by partitioning each dimension of the Iris dataset into 4, which means that the shape of V is (256,4).

2) Subtractive Clustering: The computational cost of the Mountain clustering algorithm depends on the number of nodes in the data space. To reduce this cost without compromising cluster estimations, Subtractive Clustering, a similar algorithm, is used. This method treats each data point as a potential cluster center based on its proximity to other points. Objects with higher nearby densities have greater potential values. It employs two positive radii, ra and rb, to influence the impact of surrounding points on potential cluster centers. The parameter ra defines a neighborhood for measuring potential, while rb determines the neighborhood for potential reduction.

Initially, the algorithm selects the first cluster center based on the point $x_j \in X$ with the highest density D_1, where

$$D_1(x_j) = \sum_{j=1}^{n} e^{\left(\frac{-||x_j-x_l||^2}{\frac{ra}{2}^2}\right)} \tag{3}$$

Afterwards, it starts an iterative process to find new densities D_{i+1} for each point $x_j \in X$ with the equation 4. As in the first step, the new center is defined as the point whose density is the highest.

$$D_{i+1}(x_j) = D_i(x_j) - D_i(c_i)e^{\left(\frac{-||x_j-c_i||^2}{\frac{rb}{2}^2}\right)} \tag{4}$$

This process stops once the algorithm finds a center it already found before or until it reaches the maximum number of iterations (100).

3) Naive Distance Clustering: The Naive Distance Clustering, or the "Juan's Boxes Algorithm" as we named it in class, is a clustering algorithm that uses two hyperparameters to cluster the dataset: a random point $x \in X$, and a number of clusters K.

The algorithm starts with a distance matrix D, where the entry D_{ij} is the distance between the point x_i and the point x_j. Then, it computes the maximum distance in the matrix ($maxdist$) and creates K intervals of equal size in the range [0, $maxdist$]. Finally, to assign each point to a cluster, we use the distance from said point to the random point x. If the distance falls in the first interval, then the point is assigned to the first cluster, if it falls in the last interval, it is assigned to the last cluster, and so on.

This algorithm is inherently naive since it assumes that two points with the same distance to a random point belong to the same cluster, but numerous examples contradict this assumption. Still, it is a computationally cheap algorithm to start an exploration of the dataset.

4) Connected Components Clustering: This algorithm is inspired by the concept of connected components from graph theory. A component of a given undirected graph is a connected subgraph that is not part of any larger connected subgraph. For instance, the graph shown in the first illustration has three components.

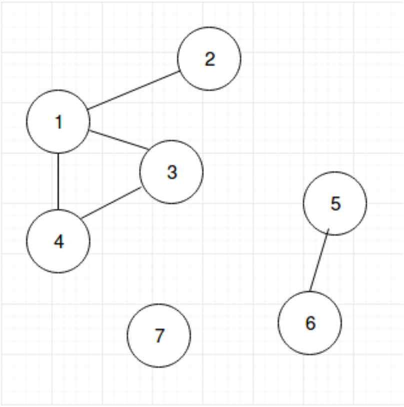

Fig. 1: Connected Components Example

For the Connected Components algorithm, we define that the points x_i and x_j are connected if their distance is less than a hyperparameter called distance threshold. Points in the same component belong to the same cluster.

The result of this clustering technique is highly dependant on the distance threshold, which is also dependant on the distance itself. As a rule, the higher the distance threshold, the fewer the clusters.

5) Fuzzy c-Means: Fuzzy c-means (FCM) is a clustering algorithm that assigns each data point to multiple clusters with varying degrees of membership, offering a more nuanced understanding of data groupings compared to traditional hard clustering methods like k-means. The key steps in FCM include:

- Initialization: Choose the number of clusters and initialize their centers randomly. Set a fuzziness parameter M, typically 2, which controls the degree of fuzziness in cluster membership.
- Membership Assignment: Compute the degree of membership for each data point in each cluster. This is done using a formula that considers the distances between data points and cluster centers, adjusted by the fuzziness parameter.
- Cluster Center Update: Update the cluster centers based on the degrees of membership and the positions of data points.
- Iteration: Repeat the membership assignment and center update steps until the cluster centers stabilize or a predefined convergence criterion is met.

FCM's output is a set of clusters where each data point has a membership degree for each cluster, indicating the strength of its association with each cluster. FCM is particularly useful in situations where cluster boundaries are not clear-cut, but it can be sensitive to initial conditions and may converge to local minima.

6) K-Means: K-means is a widely used clustering algorithm that partitions a dataset into K distinct, non-overlapping subsets, or clusters. The goal of K-means is to group data points into clusters such that the total within-cluster variance is minimized. The process starts by randomly selecting K points from the dataset as the initial centroids of the clusters. The algorithm then iteratively performs two steps until convergence: assignment and update. In the assignment step, each data point is assigned to the nearest centroid based on a distance metric, typically Euclidean distance. This forms K clusters with the data points grouped around the centroids. In the update step, the centroids of these clusters are recalculated as the mean of all data points belonging to the cluster. These steps are repeated until the centroids no longer move significantly, indicating that the clusters are as compact as possible and the data points are assigned to their nearest centroid.

Mathematically, K-means aims to minimize an objective function given by the equation 5.

$$J = \sum_{i=1}^{K}\sum_{j=1}^{n} u_{ij}||x_j - c_i||^2 \tag{5}$$

where

$$u_{ij} = \begin{cases} 1 & \text{if } ||x_j - c_i||^2 \leq ||x_j - c_l||^2; i \neq l \\ 0 & otherwise. \end{cases}$$

The simplicity of K-means makes it an efficient and popular choice for clustering in various applications. However, it also has limitations, such as sensitivity to the initial centroid placement, difficulty in identifying clusters with non-spherical shapes, and reliance on a predefined number of clusters, K.

D. *Internal Indices*

1) Silhouette Score: The silhouette score measures how similar an object is to its own cluster compared to other clusters. The value 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.

For each data point $x_i \in X$

$a(x_i)$ is the average distance from x_i to the other points in the same cluster.

$b(x_i)$ is the lowest average distance of x_i to all points in any other cluster, of which x_i is not a member.

The silhouette score for a single data point is given in equation 6:

$$s(x_i) = \frac{b(x_i) - a(x_i)}{max(a(x_i), b(x_i))} \tag{6}$$

The overall silhouette score is the average of all individual $s(x_i)$. A high silhouette score indicates that clusters are well separated and that data points are well matched to their own cluster.

2) Calinski-Harabasz (CH) Score: The CH score is a method for evaluating the dispersion of clusters. It's defined as the ratio of the sum of between-clusters dispersion and of within-cluster dispersion for all clusters. The higher the CH score, the better the clustering.

Let B be the between-group dispersion matrix, W be the within-cluster dispersion matrix, and K be the number of clusters. The CH score is computed as:

$$CHscore = \frac{Tr(B) * (n - K)}{Tr(W) * (K - 1)} \tag{7}$$

where Tr is the trace. High values generally mean better defined clusters. However, extremely high values can sometimes indicate too many clusters, particularly if the within-cluster variance is very low compared to between-cluster variance.

3) Davies-Bouldin (DB) Score: The DB score is an internal evaluation scheme for clustering algorithms where a lower DB index signifies a better clustering. It is based on the ratio of within-cluster to between-cluster distances.

For each cluster C_i:

- Compute the average distance between each point in C_i and its centroid, denoted as S_i.
- For each cluster C_j, $C_j \neq C_i$, compute the distance between the centroids of C_i and C_j, denoted as M_{ij}
- Compute $R_{ij} = \frac{S_i + S_j}{M_{ij}}$
- The DB index is given in equation 8:

$$\frac{1}{K} \sum_{i=1}^{K} max_{j \neq i} R_{ij} \tag{8}$$

E. External Indices

The external indices are used when we have true labels or external information about the data. These indices compare the clustering results with this external truth to see how well the clustering has performed, they measure things like how many data points were correctly grouped together according to the external labels.

1) Adjusted Rand Index (ARI): The Adjusted Rand Index (ARI) is a measure used to evaluate the similarity between two clusterings, often used to compare the results of a clustering algorithm with a ground truth. Unlike the raw Rand Index, which simply counts the proportion of decisions that are correct, the ARI adjusts for the chance grouping of elements, providing a more robust and meaningful assessment.

The value of the ARI ranges from -1 to 1. A score of 1 indicates perfect agreement between the two clusterings, 0 would imply random chance agreement, and negative values suggest less than chance agreement. This index is particularly valuable in assessing clustering performance in scenarios where true labels are known, allowing for a statistically sound comparison of the proposed clustering with the ground truth.

2) Normalized Mutual Information (NMI): NMI quantifies the amount of information shared between the clustering assignments and the true labels, normalized to account for the size of the clusters and the number of labels.

The formula for NMI is based on the concept of mutual information (MI), which measures the mutual dependence between two variables. For clustering, these variables are the predicted cluster assignments and the true labels. The NMI is defined as the ratio of the mutual information and the average of the entropies of each variable.

The value of NMI ranges from 0 to 1, where 0 indicates no mutual information (i.e., the clustering is independent of the true labels), and 1 signifies perfect correlation (i.e., the clustering perfectly matches the true labels). Higher NMI values indicate better clustering quality, as they reflect a higher degree of agreement between the clustering and the ground truth.

NMI is particularly effective for assessing clustering performance because it is normalized, making it less sensitive to the number of clusters or the size of the dataset. This normalization allows for a more accurate comparison of clustering results across different datasets or clustering methods.

F. Best Model Selection

To select the best models for our clustering task, we've designed a comprehensive workflow that evaluates various algorithms and their configurations across different data representations. Initially, for algorithms like connected components and naive distance clustering, we employed a grid search approach. This involved systematically varying their hyperparameters to identify configurations yielding the most favorable internal indices, which are critical in assessing the inherent clustering quality without external references.

In parallel, we took a distinct approach for the Fuzzy c-Means and K-Means algorithms. These were specifically fed with initial centers derived from mountain and subtractive clustering algorithms. Both mountain and subtractive clustering were executed under multiple hyperparameter settings to explore a range of potential initial states. This strategy aims to leverage the insights these preliminary algorithms provide, potentially enhancing the effectiveness of the Fuzzy c-Means and K-Means.

Moreover, this entire process was not limited to the original dataset. We extended our analysis to different data representations: compressed data obtained via autoencoders and UMAP (Uniform Manifold Approximation and Projection), and expanded data produced through autoencoder techniques. In particular, the compressed data has two dimensiones, while the expanded data has six. This multi-faceted approach ensures a thorough evaluation across varied data landscapes, potentially uncovering clustering patterns that might be less discernible in the original dataset.

After pinpointing the most promising models based on their internal indices, we moved to the crucial validation phase. Here, we employed external indices, which provide an objective measure of clustering quality by comparing the discovered clusters against ground truth labels. This final step is pivotal in confirming the practical applicability and reliability of our selected clustering models.

III. Results

A. Exploratory Data Analysis

1) Basic statistics: First, we computed a few simple statistics for each feature.

	SepalLength	SepalWidth	PetalLength	PetalWidth
count	150	150	150	150
mean	0.428704	0.439167	0.467571	0.457778
std	0.230018	0.180664	0.299054	0.317984
min	0.000000	0.000000	0.000000	0.000000
25%	0.222222	0.333333	0.101695	0.083333
50%	0.416667	0.416667	0.567797	0.500000
75%	0.583333	0.541667	0.694915	0.708333
max	1.000000	1.000000	1.000000	1.000000

TABLE I: Descriptive Statistics Iris Dataset

2) Histograms: Then, we delved deeper into the distribution of each feature by plotting their histograms.

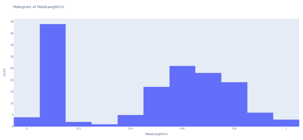

Fig. 2: Petal Length

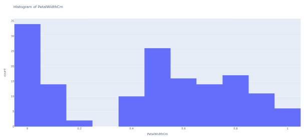

Fig. 3: Petal Width

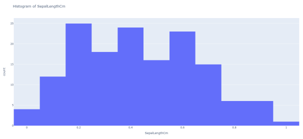

Fig. 4: Sepal Length

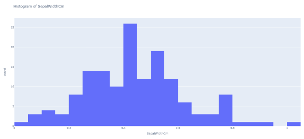

Fig. 5: Sepal Width

3) Biplot: Figure 6 displays a biplot for the data. We can clearly see two clusters.

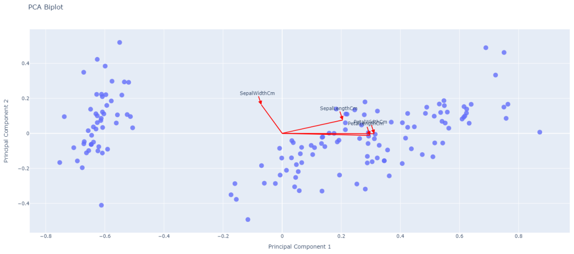

Fig. 6: Biplot

4) Pair plot: Then we took into account the labels and plotted a pair plot or a scatter matrix, shown in Figure 7.

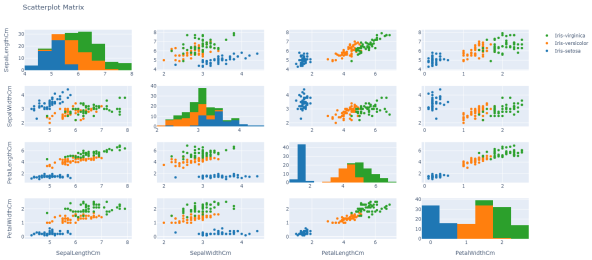

Fig. 7: Pair plot

5) UMAP: Figures 8, and 9 show the result of computing the UMAP algorithm on the Iris Dataset.

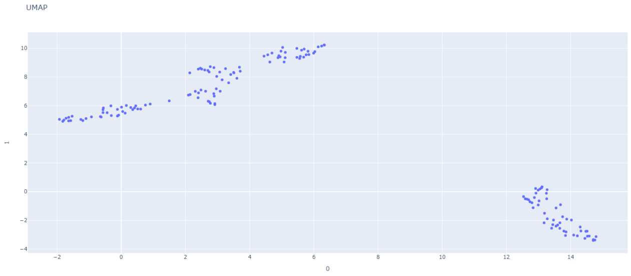

Fig. 8: UMAP unlabeled

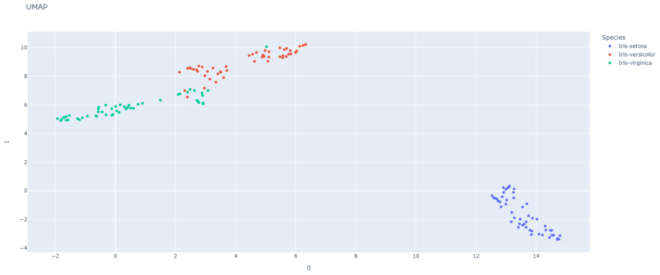

Fig. 9: UMAP labeled

B. Naive Distance Clusters

Tables II, III, IV and V show the results of the models for the different dimensions. We varied the distances and K hyperparameters, and then computed the internal indices on the resulting labels. The algorithm runs considerably fast.

Distance	K	Silhouette	DB	CH
l2	2	0.74	0.33	629
l2	3	0.61	3.40	298
l2	4	0.64	0.48	841
l2	5	0.59	0.58	961
l3	2	0.67	0.58	223
l3	3	0.62	0.54	750
l3	4	0.65	0.47	900
l3	5	0.47	3.88	202
mahal distance	2	0.51	0.92	104
mahal distance	3	0.38	4.02	26
mahal distance	4	0.41	5.18	52
mahal distance	5	0.27	5.74	21
cosine distance	2	0.84	0.36	604
cosine distance	3	0.80	0.34	323
cosine distance	4	0.76	0.62	359
cosine distance	5	0.74	0.60	231

TABLE II: Low dimensions

Distance	K	Silhouette	DB	CH
l2	2	0.40	1.87	28
l2	3	0.47	1.52	350
l2	4	0.43	3.18	182
l2	5	0.39	0.92	468
l3	2	0.42	1.77	32
l3	3	0.58	0.96	334
l3	4	0.49	0.63	540
l3	5	0.46	3.54	437
mahal distance	2	0.88	0.74	4
mahal distance	3	0.28	2.14	3
mahal distance	4	0.30	0.97	20
mahal distance	5	0.22	1.09	51
cosine distance	2	0.70	0.44	217
cosine distance	3	0.59	0.95	410
cosine distance	4	0.50	1.25	297
cosine distance	5	0.49	1.51	268

TABLE III: High dimensions

Distance	K	Silhouette	DB	CH
l2	2	0.61	0.63	287
l2	3	0.33	1.34	42
l2	4	0.40	0.89	226
l2	5	-0.04	2.27	22
l3	2	0.37	2.06	25
l3	3	0.54	0.82	287
l3	4	0.44	0.92	236
l3	5	0.22	3.31	85
mahal distance	2	0.12	6.31	3
mahal distance	3	0.04	4.43	3
mahal distance	4	0.02	8.14	12
mahal distance	5	-0.02	11.14	5
cosine distance	2	0.85	0.50	302
cosine distance	3	0.69	1.90	164
cosine distance	4	0.69	2.21	108
cosine distance	5	0.60	1.98	92

TABLE IV: Original dimensions

Distance	K	Silhouette	DB	CH
l2	2	0.77	0.38	532.37
l2	3	0.68	0.39	1506.00
l2	4	0.83	0.28	1026.00
l2	5	0.48	0.60	1401.61
l3	2	0.83	0.28	1026.00
l3	3	0.83	0.28	1026.00
l3	4	0.50	0.48	1097.10
l3	5	0.34	2.76	380.29
mahal distance	2	0.40	1.09	71.02
mahal distance	3	0.35	2.67	58.55
mahal distance	4	0.20	2.17	34.06
mahal distance	5	0.31	3.53	23.66
cosine distance	2	0.86	0.28	1026.00
cosine distance	3	0.86	0.28	1026.00
cosine distance	4	0.80	0.39	563.23
cosine distance	5	0.80	0.39	564.46

TABLE V: Low dimensions, UMAP

C. Connected Components

Tables VI, VII, VIII and IX show the results of fitting the connected components models with different distance thresholds on datasets of various dimensions. The distance thresholds for each distance were selected based on an inspection of its behaviour with the dataset. For instance, Figure 10 shows the pairwise distances in the Iris dataset. The distances range from 0 to 1.6, so we selected various thresholds accordingly. We didn't want to select a high threshold because that would create only one cluster, so we tested with the following thresholds: 0.15, 0.2, 0.25, 0.3 and 0.35. In contrast, Figure 12 shows that the maximum Mahalanobis distance was greater than 6, so we selected slightly higher thresholds, ranging between 1.2 and 1.6.

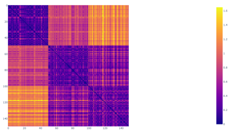

Fig. 10: l2 distances in the Iris dataset

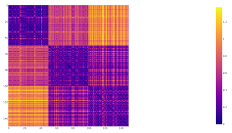

Fig. 11: l3 distances in the Iris dataset

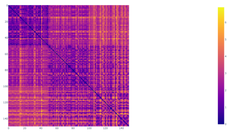

Fig. 12: Mahal distances in the Iris dataset

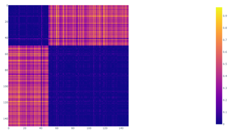

Fig. 13: Cosine distances in the Iris dataset

Distance	K	Threshold	Silhouette	DB	CH
l2	2	0.15	0.74	0.31	576
l2	2	0.2	0.74	0.31	576
l3	2	0.15	0.72	0.31	576
l3	2	0.2	0.72	0.31	576
cosine distance	3	0.01	0.81	0.35	343
cosine distance	3	0.02	0.81	0.35	343
cosine distance	2	0.03	0.88	0.31	576
cosine distance	2	0.04	0.88	0.31	576
cosine distance	2	0.05	0.88	0.31	576

TABLE VI: Low dimensions

Distance	K	Threshold	Silhouette	DB	CH
l2	4	0.15	0.33	0.43	235
l2	2	0.20	0.76	0.35	644
l2	2	0.25	0.76	0.35	644
l2	2	0.30	0.76	0.35	644
l2	2	0.35	0.76	0.35	644
l3	3	0.15	0.60	0.31	329
l3	2	0.20	0.77	0.35	644
l3	2	0.25	0.77	0.35	644
l3	2	0.30	0.77	0.35	644
mahal distance	4	1.20	0.40	0.68	3
mahal distance	4	1.30	0.40	0.68	3
mahal distance	4	1.40	0.40	0.68	3
mahal distance	4	1.50	0.40	0.68	3
mahal distance	4	1.60	0.40	0.68	3
cosine distance	3	0.01	0.66	0.31	329
cosine distance	3	0.02	0.66	0.31	329
cosine distance	3	0.03	0.66	0.31	329
cosine distance	3	0.04	0.66	0.31	329

TABLE VII: High dimensions

Distance	K	Threshold	Silhouette	DB	CH
l2	11	0.15	-0.01	0.61	53
l2	4	0.20	0.41	0.45	133
l2	3	0.25	0.56	0.43	181
l2	2	0.30	0.70	0.49	353
l2	2	0.35	0.70	0.49	353
l3	7	0.15	0.03	0.58	71
l3	3	0.20	0.55	0.43	181
l3	3	0.25	0.55	0.43	181
l3	2	0.30	0.69	0.49	353
l3	2	0.35	0.69	0.49	353
mahal distance	11	1.20	0.12	0.76	47
mahal distance	9	1.30	0.11	0.97	2
mahal distance	7	1.40	0.17	0.89	2
mahal distance	6	1.50	0.19	0.85	2
mahal distance	6	1.60	0.19	0.85	2
cosine distance	5	0.01	0.52	0.52	95
cosine distance	3	0.02	0.73	0.43	181
cosine distance	3	0.03	0.73	0.43	181
cosine distance	3	0.04	0.73	0.43	181
cosine distance	2	0.05	0.86	0.49	353

TABLE VIII: Original dimensions

Distance	K	Threshold	Silhouette	DB	CH
l2	2	0.15	0.83	0.28	1026.00
l2	2	0.20	0.83	0.28	1026.00
l2	2	0.25	0.83	0.28	1026.00
l2	2	0.30	0.83	0.28	1026.00
l2	2	0.35	0.83	0.28	1026.00
l3	2	0.15	0.83	0.28	1026.00
l3	2	0.20	0.83	0.28	1026.00
l3	2	0.25	0.83	0.28	1026.00
l3	2	0.30	0.83	0.28	1026.00
l3	2	0.35	0.83	0.28	1026.00
mahal distance	2	1.20	0.56	0.28	1026.00
mahal distance	2	1.30	0.56	0.28	1026.00
mahal distance	2	1.40	0.56	0.28	1026.00
mahal distance	2	1.50	0.56	0.28	1026.00
mahal distance	2	1.60	0.56	0.28	1026.00
cosine distance	2	0.01	0.86	0.28	1026.00
cosine distance	2	0.02	0.86	0.28	1026.00
cosine distance	2	0.03	0.86	0.28	1026.00
cosine distance	2	0.04	0.86	0.28	1026.00
cosine distance	2	0.05	0.86	0.28	1026.00

TABLE IX: Low dimensions, UMAP

D. *Fuzzy c-Means*

The Fuzzy c-Means is a soft-clustering algorithm. However, we assigned each data point to a single cluster to be able to use the internal indices from section II. To do this, we chose the cluster with the maximum degree of membership for each point.

1) Mountain Clustering: We systematically evaluated various models by adjusting the hyperparameters: σ, β, distance, and M. Due to their extensive length, the detailed results of these models are provided in the appendix (see Tables XI, XIII, and XII). These tables specifically focus on the internal indices obtained when the mountain centers are input into the algorithm. It is important to note that the cosine distance is not included in these tables. This omission is intentional; the grid used for the mountain clustering algorithm incorporates the 0 vector, rendering the cosine distance undefined in this context.

2) Subtractive Clustering: In our investigation of subtractive clustering models, we primarily focused on varying the hyperparameter ra, while maintaining rb at a constant ratio of $1.5 \times ra$. The comprehensive results from these models have been placed in the appendix due to their length. Tables XIV, XVI, and XV in the appendix highlight the internal indices corresponding to scenarios where the centers derived from subtractive clustering are fed into the algorithm.

E. *K-Means*

1) Mountain Clustering: Analogous to Fuzzy c-Means, we adjusted the hyperparameters: σ, β, distance, and M, and didn't include the cosine distances. Likewise, the results are very lengthy, so they were included in the appendix in tables XVII, XIX and XVIII.

2) Subtractive Clustering: The internal indices of the scenarios where the centers derived from subtractive clustering are fed into the K-Means algorithm are depicted in Tables XX, XXII, and XXI in the appendix.

Finally it should be noted that some tables differ in lenght in the appendix. This happens because some of the results were omitted because of indetermined divisions, inversions of singular matrices and/or the algorithms finding only one cluster.

F. Selection and validation of the best models

In our analysis, we observed that the internal indices used to evaluate our models vary significantly in their scales, rendering a direct comparison through means not only inappropriate but also potentially misleading. To address this, we adopted a ranking-based approach. Each model was ranked from best to worst based on its performance against each individual index. This method allowed us to evaluate the models relative to one another within the context of each specific index. The model that consistently appeared at the highest position across all three rankings was then selected as the most effective. This ranking methodology ensures a more equitable and accurate comparison, taking into account the unique scale and significance of each index.

After a thorough evaluation of the models generated by each algorithm, we have selected one standout model (or multiple if there's a tie) per algorithm, independent of the dimensional variations used during training. This selection was made by prioritizing the model that demonstrated the highest overall effectiveness across various internal indices, without specific consideration for the dimensionality of the training data. Moving forward, the chosen models will undergo further validation using external indices. This next phase of validation is crucial, as it provides an objective assessment of each model's performance in real-world scenarios, ensuring that our selection is not only theoretically sound but also practically viable.

1) Naive Distance Clustering: There was a tie between three models trained on UMAP, their Silhouette Scores, Davies-Bouldin indeces and Calinski-Harabasz indices were 0.83, 0.28 and 1026, respectively. The first model was trained using the l2 distance and K set to 2, while the two other models were trained using the l3 distance and K set to 2 and 3.

Their external indices were:

Model	ARI	NMI
Model 1	0.531	0.640
Model 2	0.549	0.676
Model 3	0.652	0.727

TABLE X: External Indices for Different Models

2) Connected Components Clustering: The best model was the one trained on lower dimensions, with a threshold of 0.05 and the cosine distance. This model yielded 2 clusters.

Its ARI was 0.000 and its NMI was 0.027. An ARI score of 0.000 indicates no agreement between the clustering results and the true labels beyond what would be expected by chance. In practical terms, this suggests that the model's clustering is essentially random with respect to the true data distribution. The NMI score, while slightly higher than the ARI, is still very low. A score of 0.027 indicates a very small amount of shared information between the clustering assignments and the actual labels. This suggests that the clusters formed by the model do not meaningfully correspond to the true categories in the data.

This indicates that the internal indices are highly misleading when we compute them with the cosine distance. For this reason we'll discard this distance for our next analysis.

3) Fuzzy c-Means: Ignoring the models that were fitted using cosine distances, the best model was the Fuzzy c-Means trained using the subtractive centers in low dimensions. The parameters that worked best were l2 distance, ra = 0.6, rb = 0.9, and M set to 2 and 3.

The ARI of this model was 0.0005 and the NMI was 0.0367. Similar to the former results, it seems like the model didn't cluster well.

4) K-Means: Interestingly, the best model for K-Means had the same hyperparameters and scores as the best model from Fuzzy c-Means, excepting the M hyperparameter from Fuzzy c-Means.

Its external scores are identical to the ones of the Fuzzy c-Means model.

IV. Discussion and Conclusion

In the descriptive data analyses, the medians and means exhibit a degree of similarity, although certain features display notably high standard deviations. This variance suggests that within these features, the data points are more dispersed. However, the basic statistics alone did not provide sufficient insight to solve the problem at hand. While they offered a summary of central tendencies and variations within the dataset, the presence of high standard deviations and the complexity of the underlying data distribution, as evident in the histograms and other visualizations, indicated that the data's underlying structure and relationships were not adequately captured by these basic statistics.

The histograms representing petal measurements raise suspicions as they resemble normal distributions with low kurtosis. Typically, such distributions hint at the presence of multiple populations within the dataset. A logical approach would be to initially segregate the data by petal lengths and widths before applying clustering algorithms. This approach could prove sufficient with this dataset.

Moreover, the biplot provides clear evidence of at least two distinct clusters within the dataset. Furthermore, it underscores the significance of features related to petals, as indicated by the elongated red lines on the plot.

The pair plot analysis reveals compelling insights about the Iris dataset. Notably, Iris-Setosa exhibits strikingly distinct and well-defined clusters in the joint scatter plots. Even a casual observation of the univariate feature distributions confirms the strong differences between Iris-Setosa and the other two species, Versicolor and Virginica. Conversely, Versicolor and Virginica share some similarities. However, they exhibit a linear separation in nearly all of the joint plots. This linear separation underscores the potential for using these features to differentiate between Versicolor and Virginica effectively. However, it's worth noting that exceptions exist, as the joint plots for sepal width vs. sepal length, and sepal length vs. petal width do not display this same linear separation. These nuances in the pair plot analysis offer valuable insights into the data's structure and provide a foundation for further exploration and analysis in order to distinguish between these two species with precision.

In the UMAP plot, a striking pattern emerges, confirming the presence of two distinct clusters, mirroring the insights from the earlier biplot analysis. Nevertheless, what sets this UMAP plot apart is its ability to unveil a more intricate structure within the larger cluster. In essence, it helps us see a previously hidden division within the cluster on the left, which gives rise to a total of three discernible clusters, aligning with the classifications proposed by biologists for this dataset.

In the comparative analysis of the three models based on Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) metrics for the Naive Distance Clustering, Model 3 stands out as the most effective. It achieves the highest scores in both ARI (0.652) and NMI (0.727), indicating a strong alignment with the true data distribution and a high degree of shared information between the cluster assignments and the actual labels. Model 2 shows modest improvements over Model 1, with ARI and NMI scores of 0.549 and 0.676, respectively, suggesting a better, yet still moderate, clustering effectiveness compared to Model 1, which scores 0.531 in ARI and 0.640 in NMI. Overall, while all models demonstrate a reasonable capability in clustering, Model 3's superior performance highlights its more refined or effective approach in accurately clustering the data.

As per the Connected Components Clustering, the external metrics of the selected model demonstrated that its ability to cluster was subpar. However, this is most likely a problem of the selection of the model through internal indices rather than the models themselves. We can confirm the former affirmation because, while testing the models empirically, we saw almmost perfect results with some particular distances and distance thresholds. Nevertheless, the internal indices tricked us into selecting a bad model as the best one.

Upon further inspection of the results of the Connected Components Algorithm, we discovered it had a high effectiveness in detecting outliersm, since oftentimes it created clusters with only a single data point. This characteristic makes it a valuable tool in scenarios where precise outlier identification is crucial, despite its limitations in conventional clustering.

The Mountain Clustering algorithm consistently delivered worse results than the Subtractive Clustering algorithm for both Fuzzy C-Means and K-Means. It also presented several problems, particularly in relation to cosine distances. The presence of null vectors in the grid led to issues, as this results in undefined situations in calculations based on cosines. Furthermore, we encountered difficulties with the inversion of the grid matrix when using Mahalanobis distance, which proved to be a tedious and problematic process. These challenges underscore the need for careful handling of data and metrics in such algorithms.

With the Subtractive algorithm, numerical errors emerged during computation, especially when an identified center coincided with the position of a data point in the Fuzzy C-Means method. To address this issue, it was necessary to implement certain technical adaptations and adjustments. This situation highlights the importance of anticipating and properly managing the numerical peculiarities in clustering algorithms to ensure accurate and reliable results.

Regarding the results of Fuzzy c-Means and K-Means, their performance was substantially better when they were trained independently, rather than being fed centers derived from the Mountain and Subtractive algorithms. Nonetheless, we didn't add those results to the paper.

In the context on the clusters on the Iris Dataset, we have observed a recurring tendency: there is a preference for identifying two clusters as the optimal configuration, despite the reality of the data suggesting the existence of three. This discrepancy raises questions about the effectiveness of evaluation methods based on internal indices. Additionally, it's important to highlight the limitation of cosine distance in this context; its application doesn't necessarily align with conventional metrics used in more practical and realistic evaluations. Interestingly, when the models were trained and supervised directly by me, the results were significantly better, contrasting with the less precise findings obtained using only internal indices for evaluation.

V. References

O. L. Quintero M., “Machine Intelligence for Human Decision Making,” Ed., Medellín, Colombia: Editorial Artes y Letras, 2020

Appendix

Sigma	Beta	Distance	K	M	Silhouette	DB	CH
0.50	0.50	l2	5	2	0.05	0.68	126.60
0.50	0.50	l2	5	3	0.05	0.68	126.60
0.50	1.00	l2	2	2	0.58	0.53	384.02
0.50	1.00	l2	2	3	0.58	0.53	384.02
0.50	1.50	l2	2	2	0.74	0.32	624.95
0.50	1.50	l2	2	3	0.74	0.32	624.95
1.00	0.50	l2	5	2	-0.20	1.25	47.18
1.00	0.50	l2	5	3	-0.20	1.25	47.18
1.00	1.00	l2	2	2	0.33	0.70	138.13
1.00	1.00	l2	2	3	0.33	0.70	138.13
1.00	1.50	l2	2	2	0.74	0.32	624.95
1.00	1.50	l2	2	3	0.74	0.32	624.95
1.50	0.50	l2	5	2	-0.20	1.25	47.18
1.50	0.50	l2	5	3	-0.20	1.25	47.18
1.50	1.00	l2	3	2	0.33	0.70	138.13
1.50	1.00	l2	3	3	0.33	0.70	138.13
1.50	1.50	l2	2	2	0.33	0.70	138.13
1.50	1.50	l2	2	3	0.33	0.70	138.13
0.50	0.50	l3	3	2	0.61	0.54	726.16
0.50	0.50	l3	3	3	0.61	0.54	726.16
0.50	1.00	l3	2	2	0.58	0.53	384.02
0.50	1.00	l3	2	3	0.58	0.53	384.02
1.00	0.50	l3	5	2	0.04	1.45	62.95
1.00	0.50	l3	5	3	0.04	1.45	62.95
1.00	1.00	l3	2	2	0.33	0.71	126.07
1.00	1.00	l3	2	3	0.33	0.71	126.07
1.00	1.50	l3	2	2	0.73	0.32	624.95
1.00	1.50	l3	2	3	0.73	0.32	624.95
1.50	0.50	l3	5	2	0.04	1.45	62.95
1.50	0.50	l3	5	3	0.04	1.45	62.95
1.50	1.00	l3	3	2	0.33	0.71	126.07
1.50	1.00	l3	3	3	0.33	0.71	126.07
1.50	1.50	l3	2	2	0.33	0.71	126.07
1.50	1.50	l3	2	3	0.33	0.71	126.07
0.50	0.50	mahal	8	2	0.15	1.53	112.41
0.50	0.50	mahal	8	3	0.02	2.16	147.08
0.50	1.00	mahal	4	2	0.24	1.10	291.64
0.50	1.00	mahal	3	3	0.22	4.71	63.14
0.50	1.50	mahal	2	2	0.39	0.50	296.53
0.50	1.50	mahal	2	3	0.51	0.33	472.31
1.00	0.50	mahal	15	2	0.22	0.96	465.04
1.00	0.50	mahal	3	3	0.18	3.25	33.92
1.00	1.00	mahal	5	2	0.39	0.66	139.22
1.00	1.00	mahal	5	3	0.15	1.93	85.14
1.00	1.50	mahal	3	2	0.28	1.15	294.78
1.00	1.50	mahal	3	3	0.23	1.28	258.80
1.50	0.50	mahal	11	2	0.17	1.66	134.20
1.50	0.50	mahal	15	3	0.23	0.89	589.31
1.50	1.00	mahal	6	2	0.08	6.30	19.44
1.50	1.00	mahal	6	3	0.06	1.87	23.43
1.50	1.50	mahal	3	2	0.29	0.52	321.60
1.50	1.50	mahal	3	3	0.30	0.57	338.45

TABLE XI: Fuzzy c-Means, low dimensions, mountain centers

Sigma	Beta	Distance	K	M	Silhouette	DB	CH
0.50	0.50	l2	11	2	0.58	0.62	577
0.50	0.50	l2	11	3	0.58	0.62	577
0.50	1.00	l2	2	2	0.48	0.69	248
0.50	1.00	l2	2	3	0.48	0.69	248
1.00	0.50	l2	67	2	0.14	0.64	10
1.00	0.50	l2	67	3	0.14	0.64	10
1.00	1.00	l2	6	2	0.29	0.74	105
1.00	1.00	l2	6	3	0.29	0.74	105
1.50	0.50	l2	64	2	0.02	0.73	77
1.50	0.50	l2	64	3	0.02	0.73	77
0.50	0.50	l3	5	2	0.38	0.79	461
0.50	0.50	l3	5	3	0.38	0.79	461
1.00	0.50	l3	59	2	0.19	0.62	13
1.00	0.50	l3	59	3	0.19	0.62	13
1.50	0.50	l3	61	2	0.19	0.62	13
1.50	0.50	l3	61	3	0.19	0.62	13
0.50	0.50	mahal	24	2	0.06	0.95	286
0.50	0.50	mahal	24	3	0.08	0.95	286
0.50	1.00	mahal	4	2	0.10	0.94	402
0.50	1.00	mahal	4	3	0.14	0.94	402
0.50	1.50	mahal	3	2	0.17	0.86	141
0.50	1.50	mahal	3	3	0.11	0.86	141
1.00	0.50	mahal	101	2	-0.06	1.03	266
1.00	0.50	mahal	101	3	-0.13	1.83	131
1.00	1.00	mahal	13	2	0.08	1.15	212
1.00	1.00	mahal	13	3	0.06	1.02	170
1.00	1.50	mahal	4	2	-0.03	2.30	78
1.00	1.50	mahal	4	3	-0.09	2.30	78
1.50	0.50	mahal	101	2	0.14	1.91	171
1.50	0.50	mahal	101	3	-0.11	0.96	170
1.50	1.00	mahal	101	2	0.13	2.74	208
1.50	1.00	mahal	101	3	-0.15	0.87	204
1.50	1.50	mahal	3	2	0.17	0.71	221
1.50	1.50	mahal	3	3	0.33	0.35	644

TABLE XII: Fuzzy c-Means, high dimensions, mountain centers

Sigma	Beta	Distance	K	M	Silhouette	DB	CH
0.50	0.50	l2	10	2	0.30	0.66	134.24
0.50	0.50	l2	10	3	0.30	0.66	134.24
0.50	1.00	l2	3	2	0.52	0.74	297.15
0.50	1.00	l2	3	3	0.52	0.74	297.15
0.50	1.50	l2	2	2	0.51	0.73	237.37
0.50	1.50	l2	2	3	0.51	0.73	237.37
1.00	0.50	l2	18	2	-0.14	0.93	51.42
1.00	0.50	l2	18	3	-0.14	0.93	51.42
1.00	1.00	l2	4	2	0.15	1.44	44.35
1.00	1.00	l2	4	3	0.15	1.44	44.35
1.00	1.50	l2	2	2	0.51	0.73	237.37
1.00	1.50	l2	2	3	0.51	0.73	237.37
1.50	0.50	l2	17	2	-0.10	1.13	48.40
1.50	0.50	l2	17	3	-0.10	1.13	48.40
1.50	1.00	l2	8	2	-0.27	0.88	27.80
1.50	1.00	l2	8	3	-0.27	0.88	27.80
1.50	1.50	l2	2	2	0.25	0.84	64.38
1.50	1.50	l2	2	3	0.25	0.84	64.38
0.50	0.50	l3	10	2	-0.22	0.94	37.82
0.50	0.50	l3	10	3	-0.22	0.94	37.82
0.50	1.00	l3	2	2	0.50	0.73	237.37
0.50	1.00	l3	2	3	0.50	0.73	237.37
0.50	1.50	l3	2	2	0.50	0.73	237.37
0.50	1.50	l3	2	3	0.50	0.73	237.37
1.00	0.50	l3	17	2	-0.12	0.80	39.61
1.00	0.50	l3	17	3	-0.11	0.84	39.82
1.00	1.00	l3	4	2	0.15	1.00	48.02
1.00	1.00	l3	4	3	0.15	1.00	48.04
1.00	1.50	l3	2	2	0.50	0.73	237.37
1.00	1.50	l3	2	3	0.50	0.73	237.37
1.50	0.50	l3	17	2	-0.15	1.10	35.50
1.50	0.50	l3	17	3	-0.15	1.10	35.50
1.50	1.00	l3	6	2	-0.14	1.17	30.22
1.50	1.00	l3	6	3	-0.14	1.17	30.22
1.50	1.50	l3	3	2	0.14	1.44	51.14
1.50	1.50	l3	3	3	0.14	1.44	51.14
0.50	0.50	mahal	52	2	-0.09	1.11	66.53
0.50	0.50	mahal	51	3	-0.15	1.39	52.26
0.50	1.00	mahal	4	2	0.05	1.88	49.59
0.50	1.00	mahal	4	3	0.10	1.01	233.99
0.50	1.50	mahal	2	2	0.28	0.49	353.37
0.50	1.50	mahal	2	3	0.23	0.49	353.37
1.00	0.50	mahal	101	2	-0.01	1.38	67.36
1.00	0.50	mahal	101	3	-0.04	1.25	57.05
1.00	1.00	mahal	14	2	-0.15	3.49	21.90
1.00	1.00	mahal	18	3	-0.09	1.27	125.38
1.00	1.50	mahal	4	2	0.16	2.90	177.94
1.00	1.50	mahal	4	3	0.06	3.44	40.35
1.50	0.50	mahal	101	2	-0.04	1.29	69.59
1.50	0.50	mahal	101	3	-0.03	1.16	44.59
1.50	1.00	mahal	21	2	-0.09	1.11	88.07
1.50	1.00	mahal	34	3	-0.05	0.91	94.12
1.50	1.50	mahal	6	2	0.08	1.11	86.26
1.50	1.50	mahal	10	3	0.07	1.85	114.97

TABLE XIII: Fuzzy c-Means, original dimensions, mountain centers

Ra	Rb	Distance	K	M	Silhouette	DB	CH
0.20	0.30	l2	8	2	0.44	0.70	928.31
0.20	0.30	l2	8	3	0.44	0.70	928.31
0.30	0.45	l2	7	2	0.43	0.79	709.29
0.30	0.45	l2	7	3	0.43	0.79	709.29
0.40	0.60	l2	6	2	0.46	0.77	732.22
0.40	0.60	l2	6	3	0.46	0.77	732.22
0.50	0.75	l2	5	2	0.34	0.87	567.73
0.50	0.75	l2	5	3	0.34	0.87	567.73
0.60	0.90	l2	2	2	0.74	0.32	624.95
0.60	0.90	l2	2	3	0.74	0.32	624.95
0.70	1.05	l2	3	2	0.46	0.96	319.25
0.70	1.05	l2	3	3	0.46	0.96	319.25
0.20	0.30	l3	8	2	0.48	0.67	938.34
0.20	0.30	l3	8	3	0.48	0.67	938.34
0.30	0.45	l3	6	2	0.41	0.73	709.92
0.30	0.45	l3	6	3	0.41	0.73	709.92
0.40	0.60	l3	5	2	0.44	0.84	558.29
0.40	0.60	l3	5	3	0.44	0.84	558.29
0.50	0.75	l3	5	2	0.32	0.85	314.55
0.50	0.75	l3	5	3	0.32	0.85	314.55
0.60	0.90	l3	3	2	0.46	0.90	318.00
0.60	0.90	l3	3	3	0.46	0.90	318.00
0.70	1.05	l3	3	2	0.45	0.97	319.88
0.70	1.05	l3	3	3	0.45	0.97	319.88
0.20	0.30	mahal	70	2	0.38	0.53	2072.44
0.20	0.30	mahal	71	3	0.39	0.56	1875.94
0.30	0.45	mahal	48	2	0.41	0.76	812.87
0.30	0.45	mahal	47	3	0.40	0.80	813.48
0.40	0.60	mahal	39	2	0.32	0.82	478.62
0.40	0.60	mahal	39	3	0.32	0.75	598.89
0.50	0.75	mahal	28	2	0.39	0.87	624.80
0.50	0.75	mahal	30	3	0.36	0.83	597.08
0.60	0.90	mahal	21	2	0.44	0.81	802.06
0.60	0.90	mahal	21	3	0.45	0.82	694.93
0.70	1.05	mahal	17	2	0.42	0.90	373.48
0.70	1.05	mahal	19	3	0.46	0.83	592.71
0.20	0.30	cosine	2	2	0.88	0.32	610.78
0.20	0.30	cosine	2	3	0.88	0.32	610.78
0.30	0.45	cosine	2	2	0.88	0.32	624.95
0.30	0.45	cosine	2	3	0.88	0.32	624.95
0.40	0.60	cosine	2	2	0.88	0.32	610.78
0.40	0.60	cosine	2	3	0.88	0.32	610.78
0.50	0.75	cosine	2	2	0.88	0.31	576.02
0.50	0.75	cosine	2	3	0.88	0.31	576.02
0.60	0.90	cosine	2	2	0.88	0.31	576.02
0.60	0.90	cosine	2	3	0.88	0.31	576.02
0.70	1.05	cosine	2	2	0.88	0.31	576.02
0.70	1.05	cosine	2	3	0.88	0.31	576.02

TABLE XIV: Fuzzy c-Means, low dimensions, subtractive centers

Ra	Rb	Distance	K	Silhouette	DB	CH
0.20	0.30	l2	21	0.32	0.86	466
0.20	0.30	l2	21	0.32	0.86	466
0.30	0.45	l2	9	0.35	0.78	484
0.30	0.45	l2	9	0.35	0.78	484
0.40	0.60	l2	7	0.42	0.78	567
0.40	0.60	l2	7	0.42	0.78	567
0.50	0.75	l2	6	0.43	0.78	526
0.50	0.75	l2	6	0.43	0.78	526
0.60	0.90	l2	4	0.43	1.00	384
0.60	0.90	l2	4	0.43	1.00	384
0.70	1.05	l2	5	0.38	1.03	309
0.70	1.05	l2	5	0.38	1.03	309
0.20	0.30	l3	14	0.37	0.80	474
0.20	0.30	l3	14	0.37	0.80	474
0.30	0.45	l3	7	0.39	0.79	533
0.30	0.45	l3	7	0.39	0.79	533
0.40	0.60	l3	5	0.46	0.72	565
0.40	0.60	l3	5	0.46	0.72	565
0.50	0.75	l3	4	0.43	1.00	384
0.50	0.75	l3	4	0.43	1.00	384
0.60	0.90	l3	4	0.43	0.99	373
0.60	0.90	l3	4	0.43	0.99	373
0.70	1.05	l3	4	0.42	1.00	372
0.70	1.05	l3	4	0.42	1.00	372
0.20	0.30	mahal distance	101	0.19	0.89	176
0.20	0.30	mahal distance	101	0.19	0.89	176
0.30	0.45	mahal distance	101	0.20	0.89	174
0.30	0.45	mahal distance	101	0.20	0.89	174
0.40	0.60	mahal distance	101	0.17	0.90	138
0.40	0.60	mahal distance	101	0.19	0.89	180
0.50	0.75	mahal distance	101	0.18	0.88	177
0.50	0.75	mahal distance	101	0.18	0.84	199
0.60	0.90	mahal distance	101	0.19	0.85	199
0.60	0.90	mahal distance	101	0.19	0.85	199
0.70	1.05	mahal distance	101	0.17	0.88	173
0.70	1.05	mahal distance	101	0.18	0.90	151
0.50	1.00	cosine distance	2	0.82	0.35	584
0.50	1.00	cosine distance	2	0.82	0.35	584
0.50	1.50	cosine distance	2	0.82	0.35	584
0.50	1.50	cosine distance	2	0.82	0.35	584
1.00	1.50	cosine distance	2	0.82	0.35	584
1.00	1.50	cosine distance	2	0.82	0.35	584

TABLE XV: Fuzzy c-Means, high dimensions, subtractive centers

Ra	Rb	Distance	K	M	Silhouette	DB	CH
0.20	0.30	l2	43	2	0.13	0.71	130.05
0.20	0.30	l2	43	3	0.13	0.71	130.05
0.30	0.45	l2	19	2	0.22	0.96	147.53
0.30	0.45	l2	19	3	0.22	0.96	147.53
0.40	0.60	l2	11	2	0.28	0.85	171.60
0.40	0.60	l2	11	3	0.28	0.85	171.60
0.50	0.75	l2	6	2	0.28	1.15	223.73
0.50	0.75	l2	6	3	0.28	1.15	223.73
0.60	0.90	l2	4	2	0.49	1.20	253.69
0.60	0.90	l2	4	3	0.49	1.20	253.69
0.70	1.05	l2	4	2	0.50	1.10	251.06
0.70	1.05	l2	4	3	0.50	1.10	251.06
0.20	0.30	l3	35	2	0.15	0.74	130.53
0.20	0.30	l3	35	3	0.15	0.74	130.53
0.30	0.45	l3	14	2	0.24	0.89	161.87
0.30	0.45	l3	14	3	0.24	0.89	161.87
0.40	0.60	l3	12	2	0.24	0.99	159.14
0.40	0.60	l3	12	3	0.24	0.99	159.14
0.50	0.75	l3	5	2	0.33	1.19	226.13
0.50	0.75	l3	5	3	0.33	1.19	226.13
0.60	0.90	l3	5	2	0.36	0.96	240.97
0.60	0.90	l3	5	3	0.36	0.96	240.97
0.70	1.05	l3	5	2	0.36	0.96	241.30
0.70	1.05	l3	5	3	0.36	0.96	241.30
0.20	0.30	mahal	101	2	0.15	0.78	69.84
0.20	0.30	mahal	101	3	0.17	0.74	67.02
0.30	0.45	mahal	101	2	0.16	0.67	71.87
0.30	0.45	mahal	101	3	0.17	0.65	96.74
0.40	0.60	mahal	101	2	0.16	0.64	98.55
0.40	0.60	mahal	101	3	0.17	0.65	98.99
0.50	0.75	mahal	101	2	0.16	0.65	100.46
0.50	0.75	mahal	101	3	0.16	0.66	92.00
0.60	0.90	mahal	101	2	0.13	0.69	63.07
0.60	0.90	mahal	101	3	0.13	0.69	80.75
0.70	1.05	mahal	101	2	0.11	0.71	64.66
0.70	1.05	mahal	101	3	0.13	0.70	61.77
0.20	0.30	cosine	2	2	0.86	0.49	353.37
0.20	0.30	cosine	2	3	0.86	0.49	353.37
0.30	0.45	cosine	2	2	0.86	0.49	324.36
0.30	0.45	cosine	2	3	0.86	0.49	324.36
0.40	0.60	cosine	3	2	0.62	1.44	188.93
0.40	0.60	cosine	3	3	0.62	1.44	188.93
0.50	0.75	cosine	2	2	0.86	0.49	324.36
0.50	0.75	cosine	2	3	0.86	0.49	324.36
0.60	0.90	cosine	2	2	0.86	0.49	324.36
0.60	0.90	cosine	2	3	0.86	0.49	324.36
0.70	1.05	cosine	2	2	0.60	0.96	125.21
0.70	1.05	cosine	2	3	0.60	0.96	125.21

TABLE XVI: Fuzzy c-Means, original dimensions, subtractive centers

Sigma	Beta	Distance	K	Silhouette	DB	CH
0.50	0.50	l2	5	0.05	0.68	126.60
0.50	0.50	l2	5	0.05	0.68	126.60
0.50	1.00	l2	2	0.58	0.53	384.02
0.50	1.00	l2	2	0.58	0.53	384.02
0.50	1.50	l2	2	0.74	0.32	624.95
0.50	1.50	l2	2	0.74	0.32	624.95
1.00	0.50	l2	5	-0.20	1.25	47.18
1.00	0.50	l2	5	-0.20	1.25	47.18
1.00	1.00	l2	2	0.33	0.70	138.13
1.00	1.00	l2	2	0.33	0.70	138.13
1.00	1.50	l2	2	0.74	0.32	624.95
1.00	1.50	l2	2	0.74	0.32	624.95
1.50	0.50	l2	5	-0.20	1.25	47.18
1.50	0.50	l2	5	-0.20	1.25	47.18
1.50	1.00	l2	3	0.33	0.70	138.13
1.50	1.00	l2	3	0.33	0.70	138.13
1.50	1.50	l2	2	0.33	0.70	138.13
1.50	1.50	l2	2	0.33	0.70	138.13
0.50	0.50	l3	3	0.61	0.54	726.16
0.50	0.50	l3	3	0.61	0.54	726.16
0.50	1.00	l3	2	0.58	0.53	384.02
0.50	1.00	l3	2	0.58	0.53	384.02
1.00	0.50	l3	5	0.04	1.45	62.95
1.00	0.50	l3	5	0.04	1.45	62.95
1.00	1.00	l3	2	0.33	0.71	126.07
1.00	1.00	l3	2	0.33	0.71	126.07
1.00	1.50	l3	2	0.73	0.32	624.95
1.00	1.50	l3	2	0.73	0.32	624.95
1.50	0.50	l3	5	0.04	1.45	62.95
1.50	0.50	l3	5	0.04	1.45	62.95
1.50	1.00	l3	3	0.33	0.71	126.07
1.50	1.00	l3	3	0.33	0.71	126.07
1.50	1.50	l3	2	0.33	0.71	126.07
1.50	1.50	l3	2	0.33	0.71	126.07
0.50	0.50	mahal	8	0.20	0.93	365.56
0.50	0.50	mahal	8	0.14	1.70	339.85
0.50	1.00	mahal	4	0.28	1.38	280.71
0.50	1.00	mahal	3	0.21	3.74	71.90
0.50	1.50	mahal	2	0.52	0.32	603.67
0.50	1.50	mahal	2	0.50	0.37	508.54
1.00	0.50	mahal	15	0.05	1.72	209.30
1.00	0.50	mahal	3	0.19	2.20	57.61
1.00	1.00	mahal	5	0.30	3.09	181.37
1.00	1.00	mahal	5	0.18	2.76	90.85
1.00	1.50	mahal	3	0.16	1.13	131.63
1.00	1.50	mahal	3	0.31	1.49	301.12
1.50	0.50	mahal	11	0.05	1.48	132.51
1.50	0.50	mahal	15	-0.02	1.72	205.27
1.50	1.00	mahal	6	0.02	6.53	31.71
1.50	1.00	mahal	6	0.10	3.39	48.38
1.50	1.50	mahal	3	0.29	0.57	338.45
1.50	1.50	mahal	3	0.44	0.79	183.98

TABLE XVII: K-Means, low dimensions, mountain centers

Sigma	Beta	Distance	K	Silhouette	DB	CH
0.50	0.50	l2	11	0.58	0.62	577
0.50	0.50	l2	11	0.58	0.62	577
0.50	1.00	l2	2	0.48	0.69	248
0.50	1.00	l2	2	0.48	0.69	248
1.00	0.50	l2	67	0.14	0.64	10
1.00	0.50	l2	67	0.14	0.64	10
1.00	1.00	l2	6	0.29	0.74	105
1.00	1.00	l2	6	0.29	0.74	105
1.50	0.50	l2	64	0.02	0.73	77
1.50	0.50	l2	64	0.02	0.73	77
0.50	0.50	l3	5	0.38	0.79	461
0.50	0.50	l3	5	0.38	0.79	461
1.00	0.50	l3	59	0.19	0.62	13
1.00	0.50	l3	59	0.19	0.62	13
1.50	0.50	l3	61	0.19	0.62	13
1.50	0.50	l3	61	0.19	0.62	13
0.50	0.50	mahal	24	0.08	0.95	286
0.50	0.50	mahal	24	-0.02	1.70	102
0.50	1.00	mahal	4	0.11	0.94	402
0.50	1.00	mahal	4	0.09	0.94	402
0.50	1.50	mahal	3	0.17	0.86	141
0.50	1.50	mahal	3	0.07	0.86	141
1.00	0.50	mahal	101	-0.06	1.03	266
1.00	0.50	mahal	101	-0.04	1.03	266
1.00	1.00	mahal	13	-0.01	1.02	170
1.00	1.00	mahal	13	0.06	1.02	170
1.00	1.50	mahal	4	-0.03	2.30	78
1.00	1.50	mahal	4	-0.03	2.30	78
1.50	0.50	mahal	101	-0.05	1.82	202
1.50	0.50	mahal	101	-0.10	0.78	284
1.50	1.00	mahal	101	0.13	1.94	256
1.50	1.00	mahal	101	0.12	0.58	362
1.50	1.50	mahal	3	0.17	0.71	221
1.50	1.50	mahal	3	0.33	0.35	644

TABLE XVIII: K-Means, high dimensions, mountain centers

Sigma	Beta	Distance	K	Silhouette	DB	CH
0.50	0.50	l2	10	0.30	0.66	134.24
0.50	0.50	l2	10	0.30	0.66	134.24
0.50	1.00	l2	3	0.52	0.74	297.15
0.50	1.00	l2	3	0.52	0.74	297.15
0.50	1.50	l2	2	0.51	0.73	237.37
0.50	1.50	l2	2	0.51	0.73	237.37
1.00	0.50	l2	18	-0.14	0.93	51.42
1.00	0.50	l2	18	-0.14	0.93	51.42
1.00	1.00	l2	4	0.15	1.44	44.35
1.00	1.00	l2	4	0.15	1.44	44.35
1.00	1.50	l2	2	0.51	0.73	237.37
1.00	1.50	l2	2	0.51	0.73	237.37
1.50	0.50	l2	17	-0.10	1.13	48.40
1.50	0.50	l2	17	-0.10	1.13	48.40
1.50	1.00	l2	8	-0.27	0.88	27.80
1.50	1.00	l2	8	-0.27	0.88	27.80
1.50	1.50	l2	2	0.25	0.84	64.38
1.50	1.50	l2	2	0.25	0.84	64.38
0.50	0.50	l3	10	-0.22	0.94	37.82
0.50	0.50	l3	10	-0.22	0.94	37.82
0.50	1.00	l3	2	0.50	0.73	237.37
0.50	1.00	l3	2	0.50	0.73	237.37
0.50	1.50	l3	2	0.50	0.73	237.37
0.50	1.50	l3	2	0.50	0.73	237.37
1.00	0.50	l3	17	-0.12	0.80	39.61
1.00	0.50	l3	17	-0.12	0.80	39.61
1.00	1.00	l3	4	0.15	1.00	48.02
1.00	1.00	l3	4	0.15	1.00	48.02
1.00	1.50	l3	2	0.50	0.73	237.37
1.00	1.50	l3	2	0.50	0.73	237.37
1.50	0.50	l3	17	-0.15	1.10	35.50
1.50	0.50	l3	17	-0.15	1.10	35.50
1.50	1.00	l3	6	-0.14	1.17	30.22
1.50	1.00	l3	6	-0.14	1.17	30.22
1.50	1.50	l3	3	0.14	1.44	51.14
1.50	1.50	l3	3	0.14	1.44	51.14
0.50	0.50	mahal	52	-0.04	3.75	76.02
0.50	0.50	mahal	51	-0.17	1.74	76.18
0.50	1.00	mahal	4	0.08	1.37	171.63
0.50	1.00	mahal	4	0.06	1.50	153.69
0.50	1.50	mahal	2	0.05	3.04	13.23
0.50	1.50	mahal	2	0.34	0.49	324.36
1.00	0.50	mahal	101	-0.21	2.12	35.88
1.00	0.50	mahal	101	-0.18	1.26	59.81
1.00	1.00	mahal	14	-0.10	1.40	81.33
1.00	1.00	mahal	18	0.15	1.07	116.77
1.00	1.50	mahal	4	0.07	1.35	168.04
1.00	1.50	mahal	4	0.04	1.39	170.28
1.50	0.50	mahal	101	-0.04	2.98	61.25
1.50	0.50	mahal	101	0.06	4.25	94.45
1.50	1.00	mahal	21	-0.06	1.65	103.36
1.50	1.00	mahal	34	-0.08	1.10	26.79
1.50	1.50	mahal	6	0.05	1.11	86.26
1.50	1.50	mahal	10	0.07	1.85	114.97

TABLE XIX: K-Means, original dimensions, mountain centers

Ra	Rb	Distance	K	Silhouette	DB	CH
0.20	0.30	l2	8	0.44	0.70	928.31
0.20	0.30	l2	8	0.44	0.70	928.31
0.30	0.45	l2	7	0.43	0.79	709.29
0.30	0.45	l2	7	0.43	0.79	709.29
0.40	0.60	l2	6	0.46	0.77	732.22
0.40	0.60	l2	6	0.46	0.77	732.22
0.50	0.75	l2	5	0.34	0.87	567.73
0.50	0.75	l2	5	0.34	0.87	567.73
0.60	0.90	l2	2	0.74	0.32	624.95
0.60	0.90	l2	2	0.74	0.32	624.95
0.70	1.05	l2	3	0.46	0.96	319.25
0.70	1.05	l2	3	0.46	0.96	319.25
0.20	0.30	l3	8	0.48	0.67	938.34
0.20	0.30	l3	8	0.48	0.67	938.34
0.30	0.45	l3	6	0.41	0.73	709.92
0.30	0.45	l3	6	0.41	0.73	709.92
0.40	0.60	l3	5	0.44	0.84	558.29
0.40	0.60	l3	5	0.44	0.84	558.29
0.50	0.75	l3	5	0.32	0.85	314.55
0.50	0.75	l3	5	0.32	0.85	314.55
0.60	0.90	l3	3	0.46	0.90	318.00
0.60	0.90	l3	3	0.46	0.90	318.00
0.70	1.05	l3	3	0.45	0.97	319.88
0.70	1.05	l3	3	0.45	0.97	319.88
0.20	0.30	mahal	70	0.37	0.43	2267.89
0.20	0.30	mahal	71	0.39	0.48	2075.13
0.30	0.45	mahal	48	0.38	0.70	842.76
0.30	0.45	mahal	47	0.39	0.80	825.73
0.40	0.60	mahal	39	0.32	0.79	488.98
0.40	0.60	mahal	39	0.35	0.76	580.43
0.50	0.75	mahal	28	0.33	0.85	607.64
0.50	0.75	mahal	30	0.34	0.89	592.65
0.60	0.90	mahal	21	0.35	0.81	770.69
0.60	0.90	mahal	21	0.33	0.81	753.14
0.70	1.05	mahal	17	0.35	0.94	438.15
0.70	1.05	mahal	19	0.33	0.79	682.20
0.20	0.30	cosine	2	0.88	0.32	610.78
0.20	0.30	cosine	2	0.88	0.32	610.78
0.30	0.45	cosine	2	0.88	0.32	624.95
0.30	0.45	cosine	2	0.88	0.32	624.95
0.40	0.60	cosine	2	0.88	0.32	610.78
0.40	0.60	cosine	2	0.88	0.32	610.78
0.50	0.75	cosine	2	0.88	0.31	576.02
0.50	0.75	cosine	2	0.88	0.31	576.02
0.60	0.90	cosine	2	0.88	0.31	576.02
0.60	0.90	cosine	2	0.88	0.31	576.02
0.70	1.05	cosine	2	0.88	0.31	576.02
0.70	1.05	cosine	2	0.88	0.31	576.02

TABLE XX: K-Means, low dimensions, subtractive centers

Ra	Rb	Distance	K	Silhouette	DB	CH
0.20	0.30	l2	21	0.32	0.86	466.07
0.20	0.30	l2	21	0.32	0.86	466.07
0.30	0.45	l2	9	0.35	0.78	484.47
0.30	0.45	l2	9	0.35	0.78	484.47
0.40	0.60	l2	7	0.42	0.78	566.71
0.40	0.60	l2	7	0.42	0.78	566.71
0.50	0.75	l2	6	0.43	0.78	525.89
0.50	0.75	l2	6	0.43	0.78	525.89
0.60	0.90	l2	4	0.43	1.00	383.91
0.60	0.90	l2	4	0.43	1.00	383.91
0.70	1.05	l2	5	0.38	1.03	309.21
0.70	1.05	l2	5	0.38	1.03	309.21
0.20	0.30	l3	14	0.37	0.80	474.30
0.20	0.30	l3	14	0.37	0.80	474.30
0.30	0.45	l3	7	0.39	0.79	533.22
0.30	0.45	l3	7	0.39	0.79	533.22
0.40	0.60	l3	5	0.46	0.72	565.47
0.40	0.60	l3	5	0.46	0.72	565.47
0.50	0.75	l3	4	0.43	1.00	383.91
0.50	0.75	l3	4	0.43	1.00	383.91
0.60	0.90	l3	4	0.43	0.99	373.03
0.60	0.90	l3	4	0.43	0.99	373.03
0.70	1.05	l3	4	0.42	1.00	371.72
0.70	1.05	l3	4	0.42	1.00	371.72
0.20	0.30	mahal	101	0.19	0.89	175.74
0.20	0.30	mahal	101	0.19	0.89	175.74
0.30	0.45	mahal	101	0.20	0.89	174.49
0.30	0.45	mahal	101	0.20	0.89	174.49
0.40	0.60	mahal	101	0.17	0.90	138.25
0.40	0.60	mahal	101	0.19	0.89	180.24
0.50	0.75	mahal	101	0.18	0.88	176.76
0.50	0.75	mahal	101	0.18	0.84	199.31
0.60	0.90	mahal	101	0.19	0.84	199.35
0.60	0.90	mahal	101	0.19	0.84	199.35
0.70	1.05	mahal	101	0.18	0.88	151.41
0.70	1.05	mahal	101	0.18	0.88	151.41

TABLE XXI: K-Means, high dimensions, subtractive centers

Ra	Rb	Distance	K	Silhouette	DB	CH
0.20	0.30	l2	43	0.13	0.71	130.05
0.20	0.30	l2	43	0.13	0.71	130.05
0.30	0.45	l2	19	0.22	0.96	147.53
0.30	0.45	l2	19	0.22	0.96	147.53
0.40	0.60	l2	11	0.28	0.85	171.60
0.40	0.60	l2	11	0.28	0.85	171.60
0.50	0.75	l2	6	0.28	1.15	223.73
0.50	0.75	l2	6	0.28	1.15	223.73
0.60	0.90	l2	4	0.49	1.20	253.69
0.60	0.90	l2	4	0.49	1.20	253.69
0.70	1.05	l2	4	0.50	1.10	251.06
0.70	1.05	l2	4	0.50	1.10	251.06
0.20	0.30	l3	35	0.15	0.74	130.53
0.20	0.30	l3	35	0.15	0.74	130.53
0.30	0.45	l3	14	0.24	0.89	161.87
0.30	0.45	l3	14	0.24	0.89	161.87
0.40	0.60	l3	12	0.24	0.99	159.14
0.40	0.60	l3	12	0.24	0.99	159.14
0.50	0.75	l3	5	0.33	1.19	226.13
0.50	0.75	l3	5	0.33	1.19	226.13
0.60	0.90	l3	5	0.36	0.96	240.97
0.60	0.90	l3	5	0.36	0.96	240.97
0.70	1.05	l3	5	0.36	0.96	241.30
0.70	1.05	l3	5	0.36	0.96	241.30
0.20	0.30	mahal	101	0.13	0.78	71.17
0.20	0.30	mahal	101	0.17	0.70	77.94
0.30	0.45	mahal	101	0.16	0.67	84.68
0.30	0.45	mahal	101	0.15	0.63	97.17
0.40	0.60	mahal	101	0.16	0.66	92.73
0.40	0.60	mahal	101	0.17	0.66	98.57
0.50	0.75	mahal	101	0.16	0.65	100.53
0.50	0.75	mahal	101	0.16	0.66	92.00
0.60	0.90	mahal	101	0.13	0.69	75.74
0.60	0.90	mahal	101	0.13	0.69	80.75
0.70	1.05	mahal	101	0.14	0.71	66.57
0.70	1.05	mahal	101	0.12	0.70	61.77
0.20	0.30	cosine	2	0.86	0.49	353.37
0.20	0.30	cosine	2	0.86	0.49	353.37
0.30	0.45	cosine	2	0.86	0.49	324.36
0.30	0.45	cosine	2	0.86	0.49	324.36
0.40	0.60	cosine	3	0.62	1.44	188.93
0.40	0.60	cosine	3	0.62	1.44	188.93
0.50	0.75	cosine	2	0.86	0.49	324.36
0.50	0.75	cosine	2	0.86	0.49	324.36
0.60	0.90	cosine	2	0.86	0.49	324.36
0.60	0.90	cosine	2	0.86	0.49	324.36
0.70	1.05	cosine	2	0.60	0.96	125.21
0.70	1.05	cosine	2	0.60	0.96	125.21

TABLE XXII: K-Means, original dimensions, subtractive centers