LUIGI VANVITELLI UNIVERSITY
MATHEMATICS AND PHYSICS

march 23, 2024

NUMERICAL METHODS ASSIGNMENTS

by

Bamdad Booyeh

1 Question 1 (Iris Dataset)

Load from Blackboard the data file IrisDataAnnotated mat. The data set X is the same as the previously used one, while the annotation vector I indicates the iris species,

1 = Iris setosa, 2 = Iris versacolor, 3 = Iris virginica.

(a) Write your own k-means and k-medoids algorithms. In your k-medoid algorithm, use the l-distance,

(b) Run the k-means and k-medoids algorithms with k = 3. Use a random initialization to avoid putting the data by chance in three correct groups.

If the data were a suitable target for the algorithm, you should have the three species in three different clusters. Investigate the quality of the clustering by using the annotation vector of the data: Decide which iris species in each of your cluster represents by a majority, then count the misclassifications of each iris type. Run the test a couple of times to see that the result is not too sensitive to initial clustering, or, if it turns out to be, report that finding.

1.1 Answer:

Using the k-means and k-medoids algorithms, we found that both can work well but not flawlessly. The k-medoids algorithm performed slightly better for this dataset. We implemented k-medoids with both norm 1 (Manhattan distance) and norm 2 (Euclidean distance). Both norms achieved the same accuracy.

For k-medoids with norm 1, the selected medoids were samples number 8, 27, and 129. The confusion matrix showed that only 13 out of the 150 samples were misclassified. This was the same accuracy as with norm 2. However, this accuracy was not good enough, so we applied Linear Discriminant Analysis (LDA).

LDA works by maximizing the distance between different classes while minimizing the distance within each class. Using LDA, we achieved much better results, with only three misclassified individual flowers. We also compared the accuracy of Principal Component Analysis (PCA) with LDA. For this specific dataset, LDA clearly performed much better than PCA.

In summary, while k-means and k-medoids both provided reasonable clustering results, LDA significantly improved the clustering accuracy. This shows that for the Iris dataset, LDA is more effective in distinguishing between the three species.

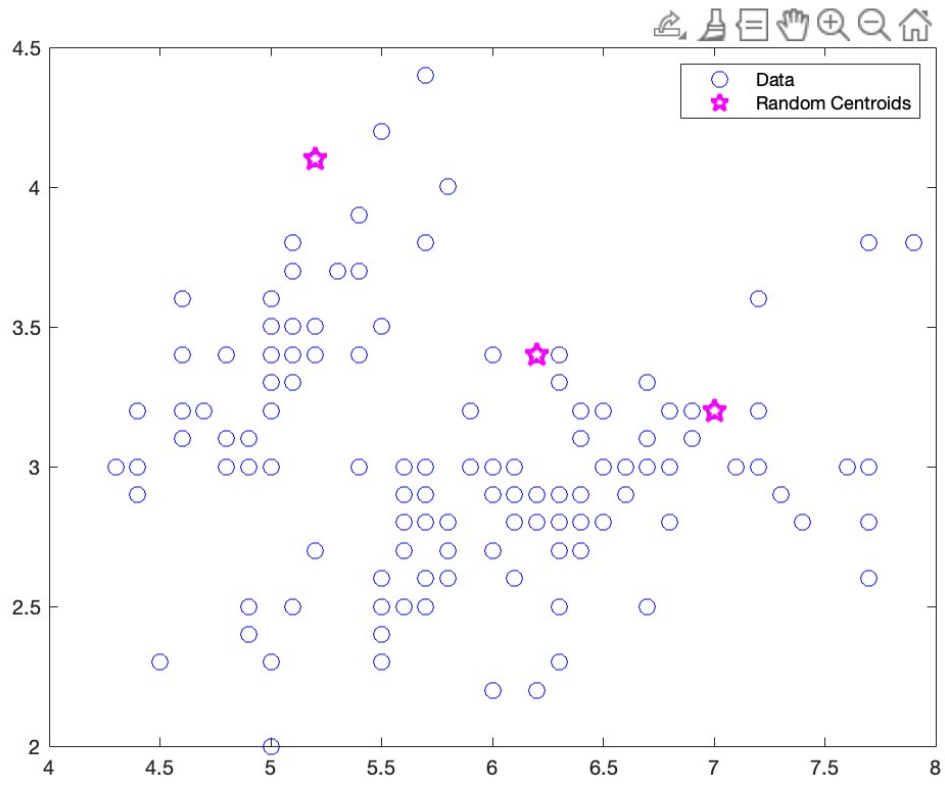

Abbildung 1: Random Medoids

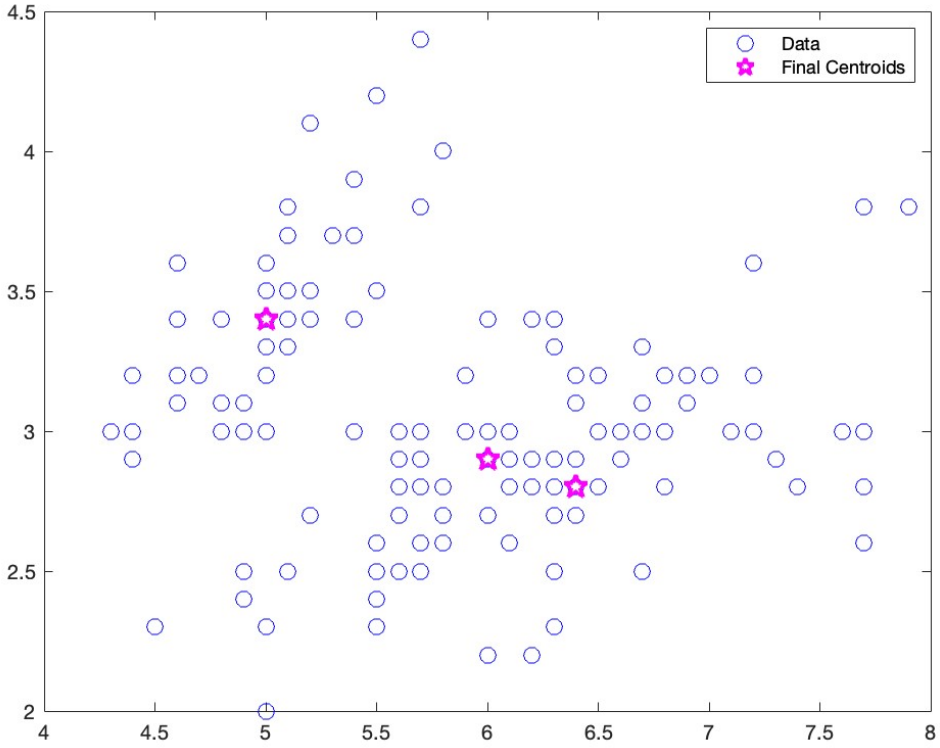

Abbildung 2: Final Medoids Norm1

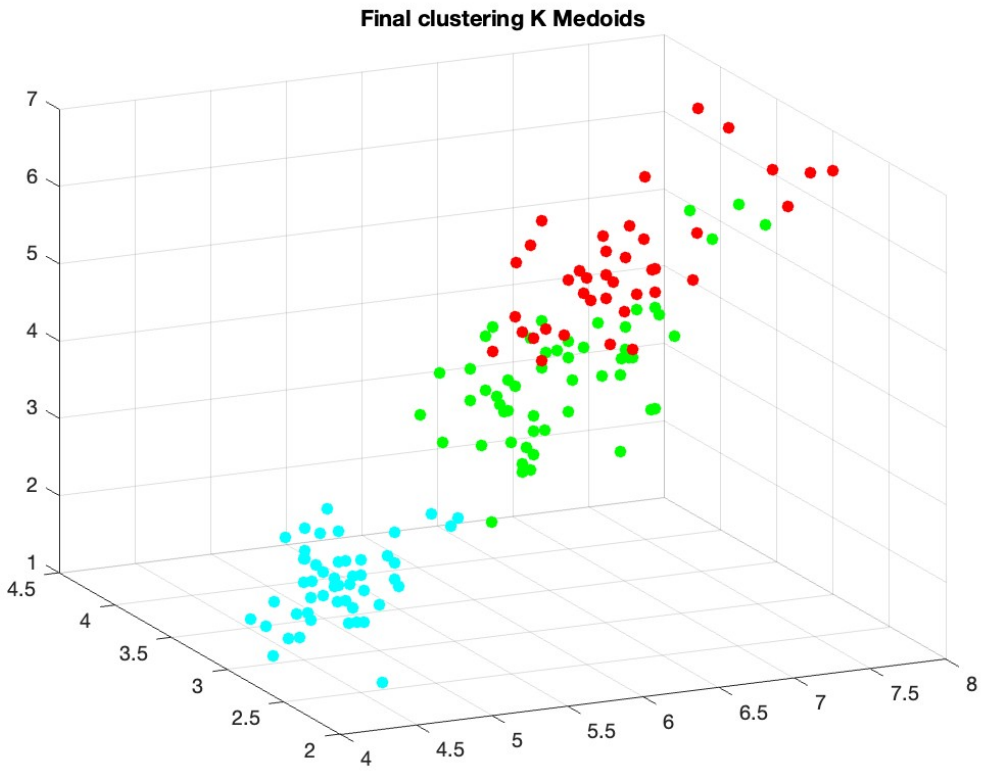

Abbildung 3: Final Clusters K-Medoids Norm1

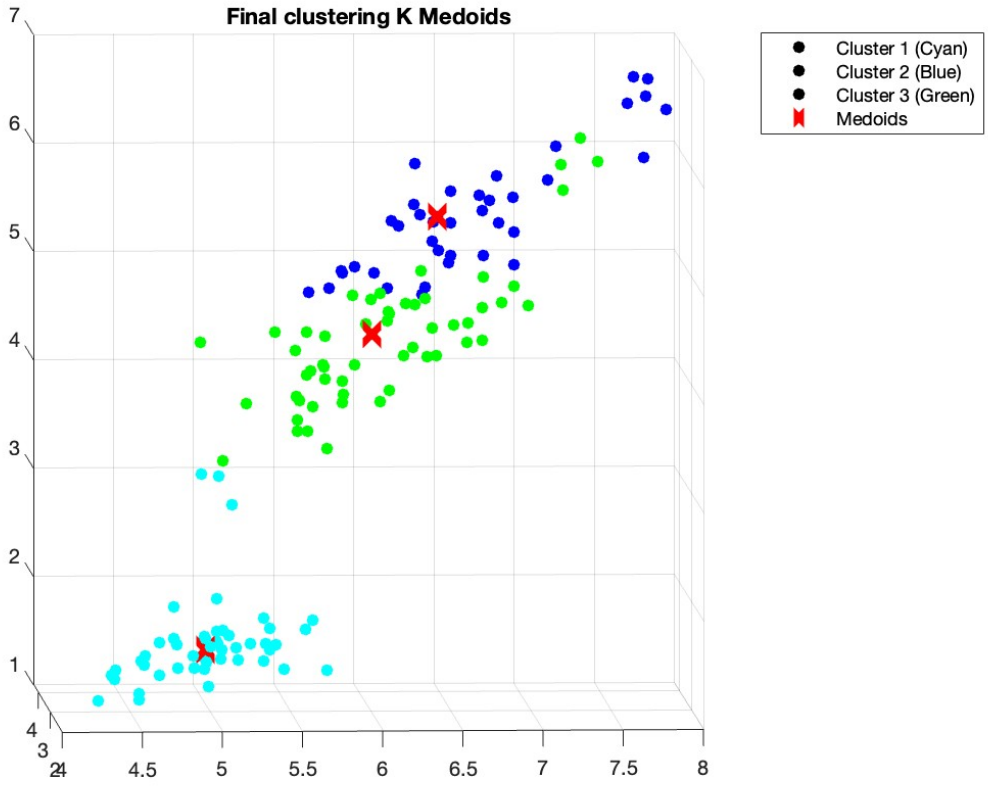

Abbildung 4: Final Clusters With Medoids Norm1

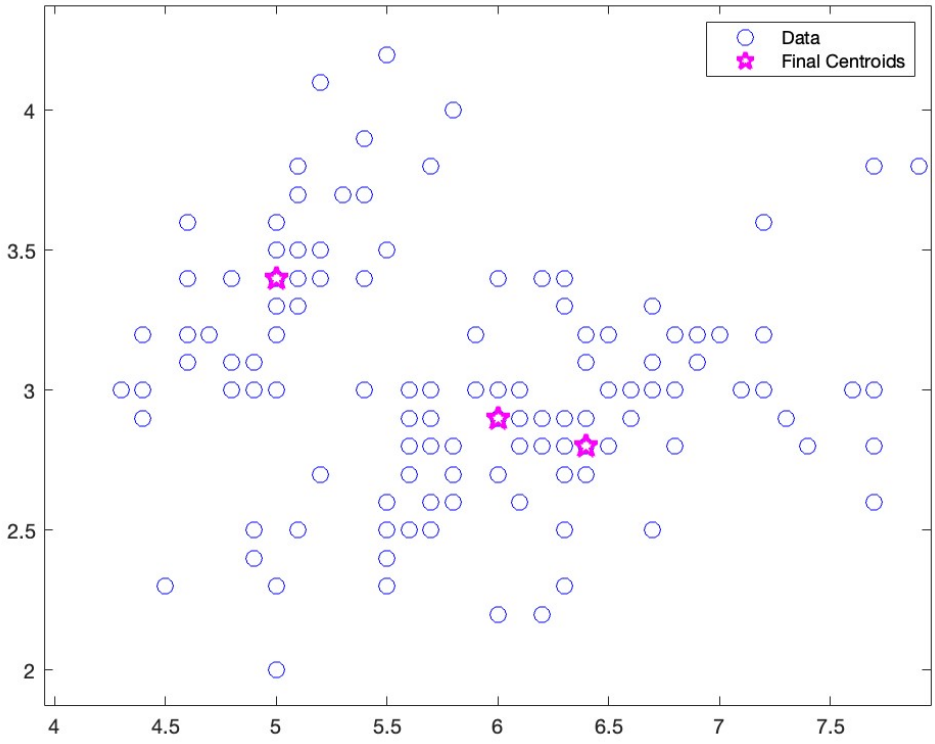

Abbildung 5: Final Medoids Norm2

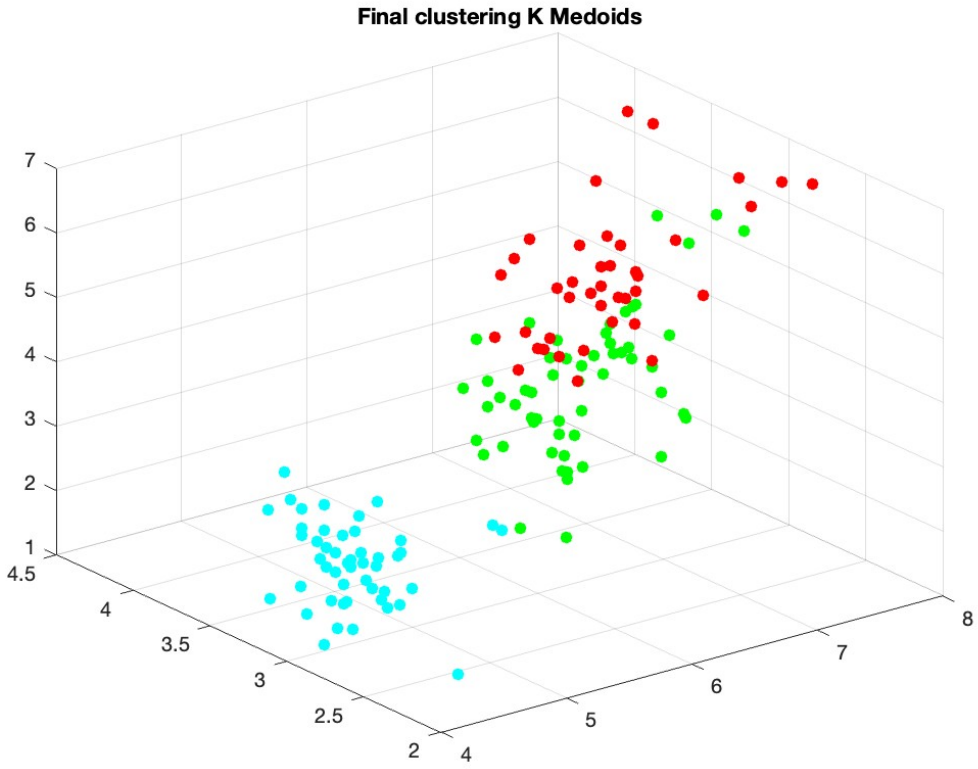

Abbildung 6: Final Clusters Norm2

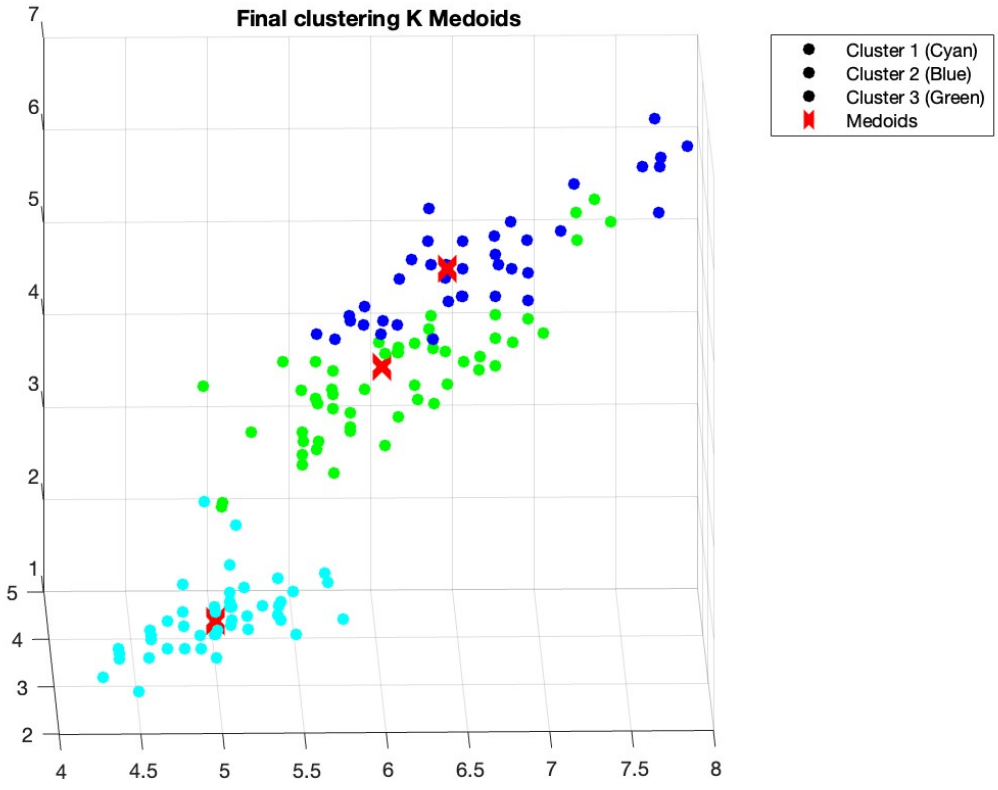

Abbildung 7: Final Clusters With Medoids Norm2

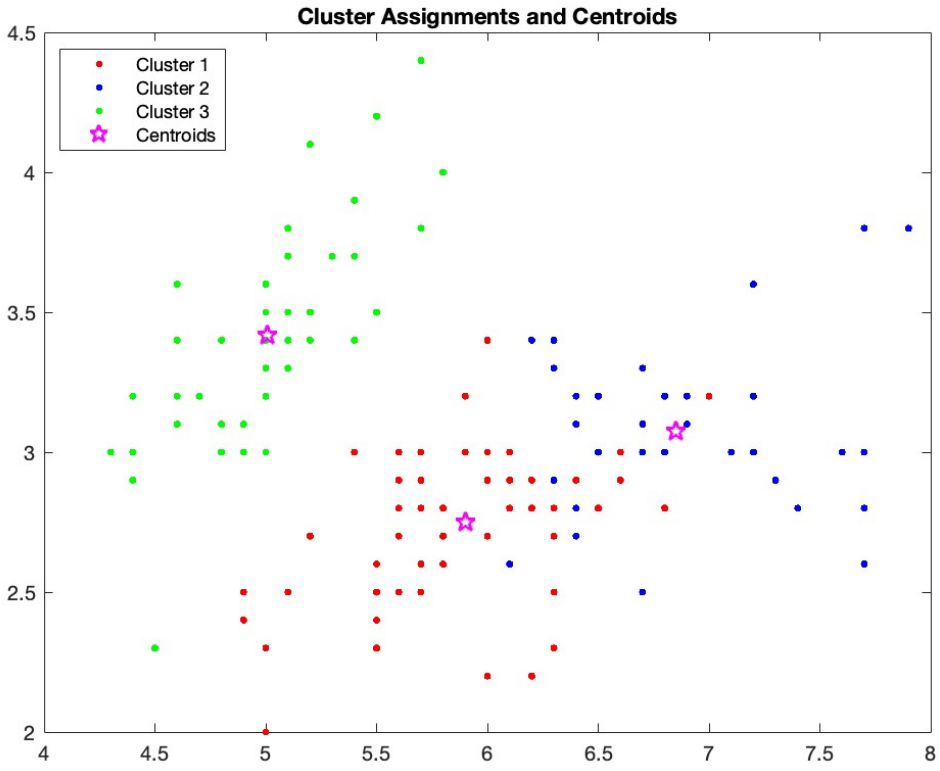

Abbildung 8: Centroids Using K-Mean

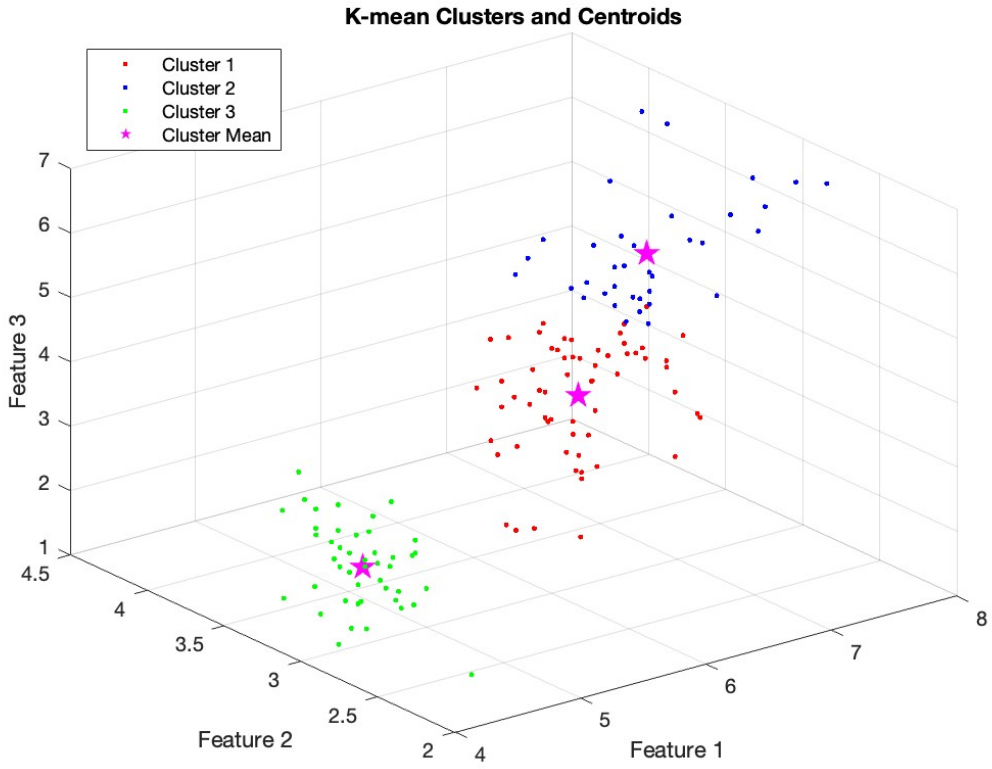

Abbildung 9: Final Clusters Using K-Mean

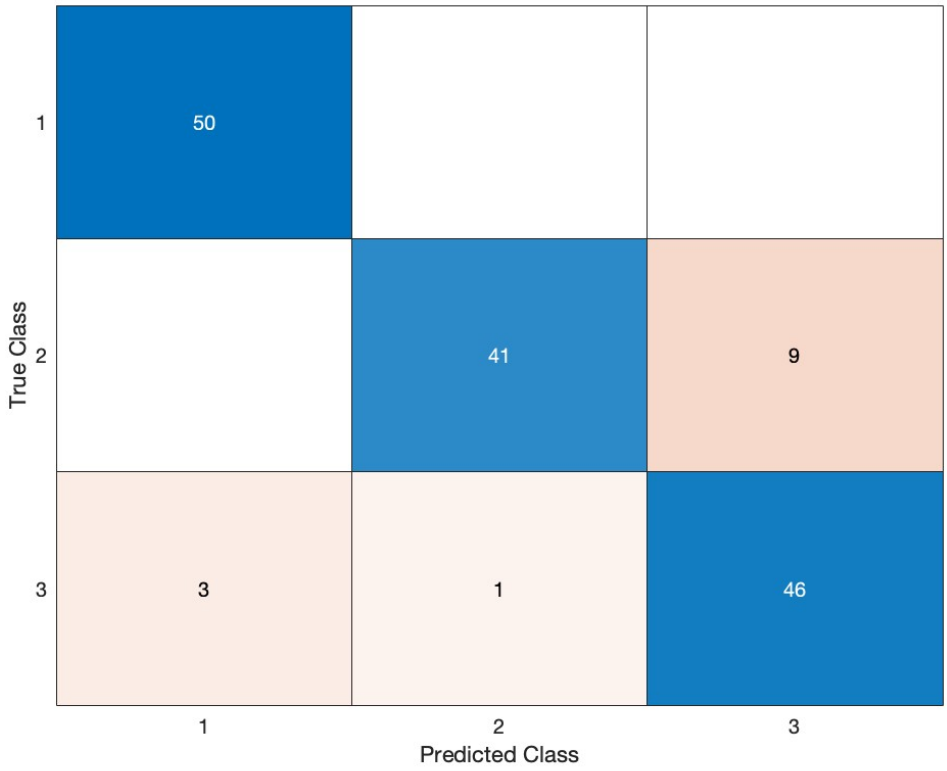

Abbildung 10: Confusion Chart Using K-Medoid Norm1

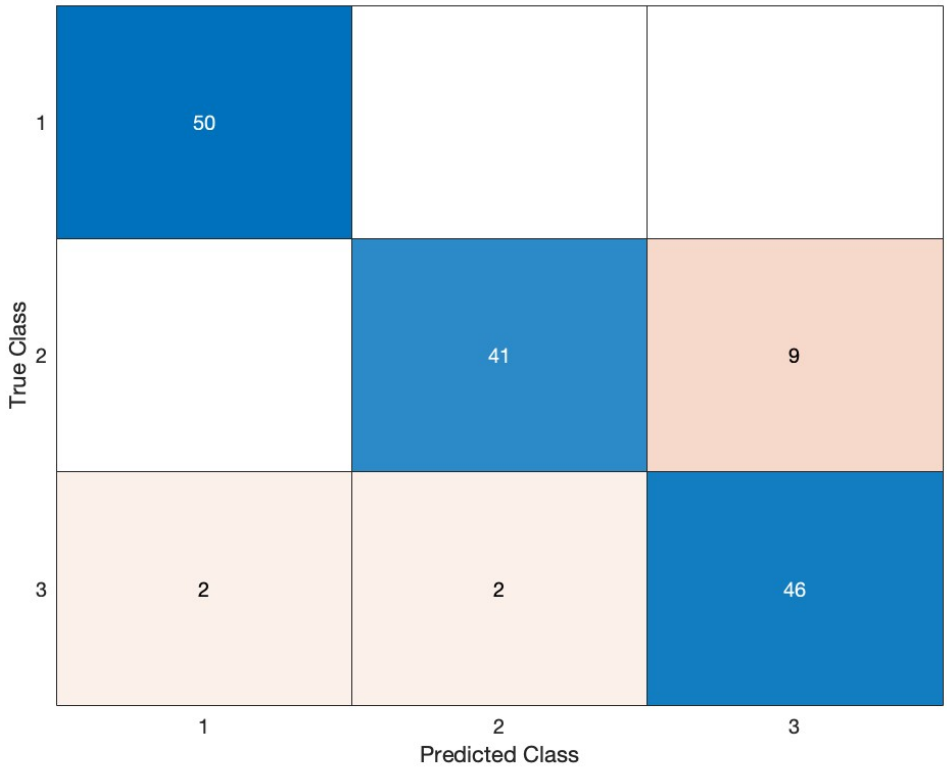

Abbildung 11: Confusion Chart Using K-Medoid Norm1

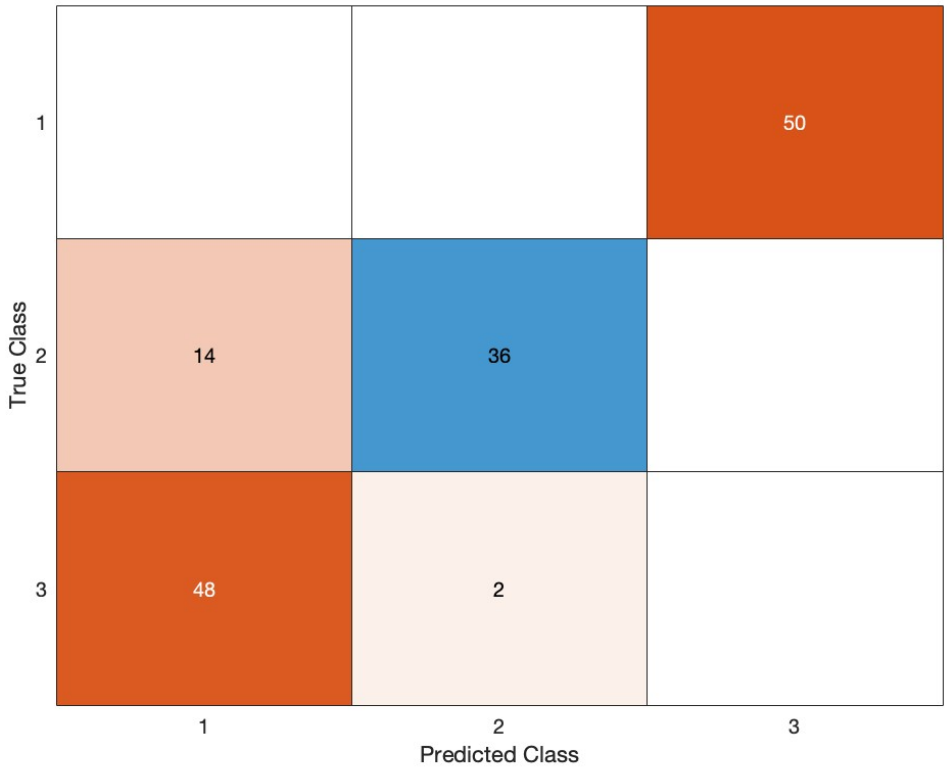

Abbildung 12: Confusion Chart Using K-Mean

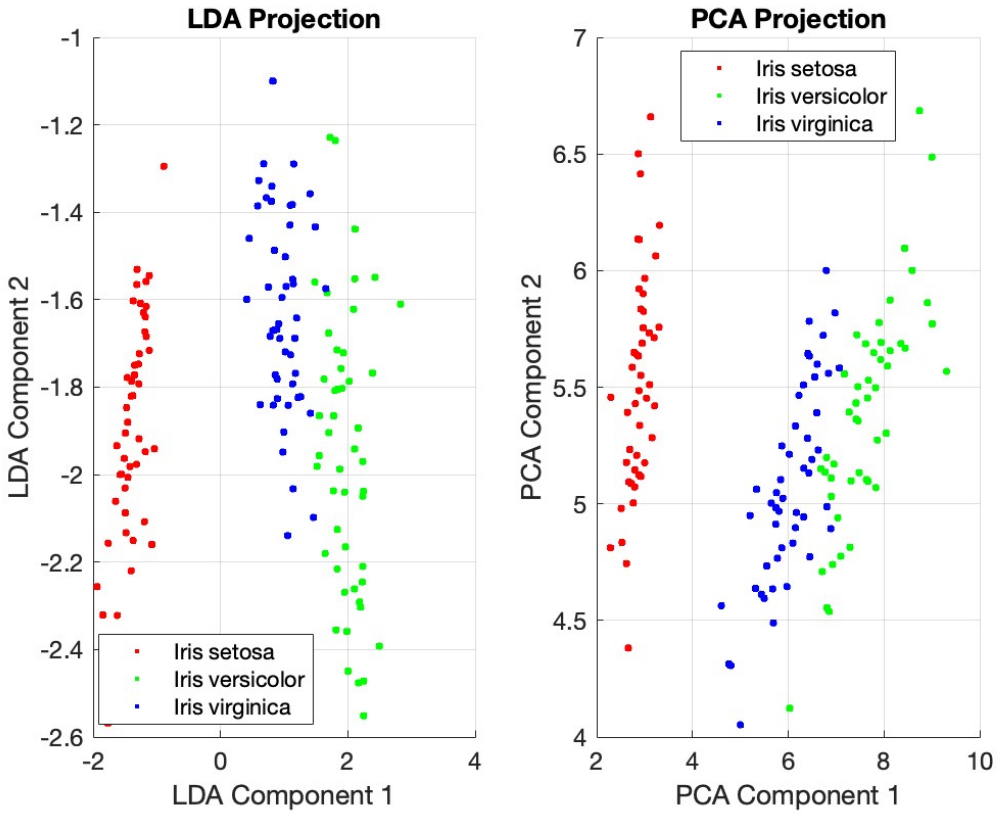

Abbildung 13: LDA vs PCA Irisdata

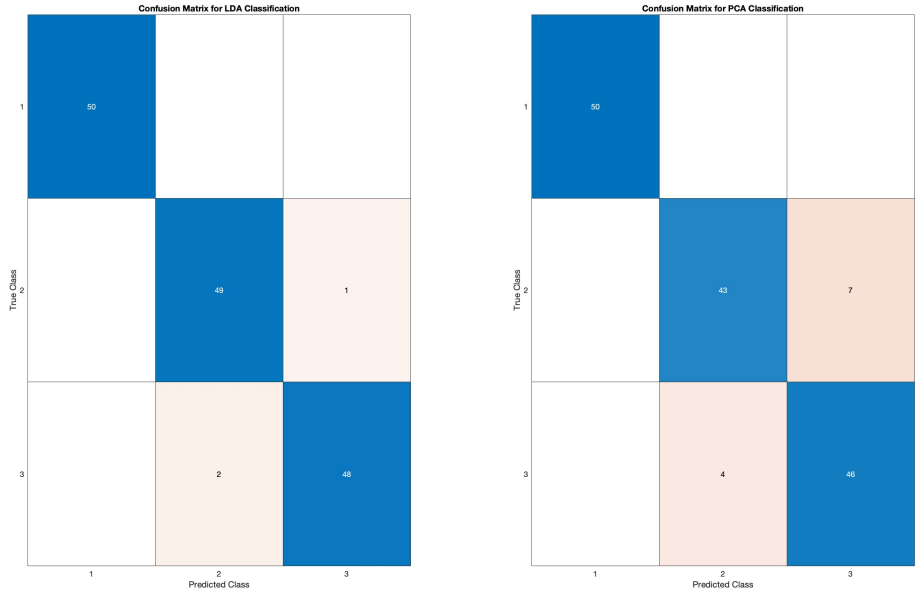

Abbildung 14: Confusion Chart LDA vs PCA

2 Question 2(Biopsy Dataset)

Download from Blackboard the file BiopsyDataAnnotated.mat. The file contains a data matrix X of size 9 × 699, containing breast tissue needle biopsy data from 699 patients, some of which have breast cancer, and some have a benign tumor. The explanation of the columns is as follows:

Each attribute takes on a value between 1 and 10. Some data is missing, which is indicated by a 'NaN' (= not a number) in the file. The vector I is an annotation vector, with the annotation

1 = malignant,

0 = benign.

After deleting columns containing missing data, run your k-medoids algorithm. Then check if the clustering corresponds to the annotation. Investigate the success of the classifier in terms of misclassification. Calculate the specificity (or true negative rate) and sensitivity (or recall rate) of the method. Again, run your algorithm a couple of times to assess the robustness of the algorithm to initial partitioning.

2.1 Answer:

The k-medoids algorithm with both Manhattan and Euclidean distances performed reasonably well on the Biopsy dataset after handling missing data. However, due to the absence of the true annotation vector during the analysis, we relied on visualization to assess clustering quality. To accurately evaluate the clustering performance, it is essential to compare the algorithm's output with the true annotations and calculate misclassification rate, sensitivity, and specificity. The algorithm's sensitivity to initial partitioning suggests that multiple runs and averaging results could provide more stable and reliable clustering outcomes.

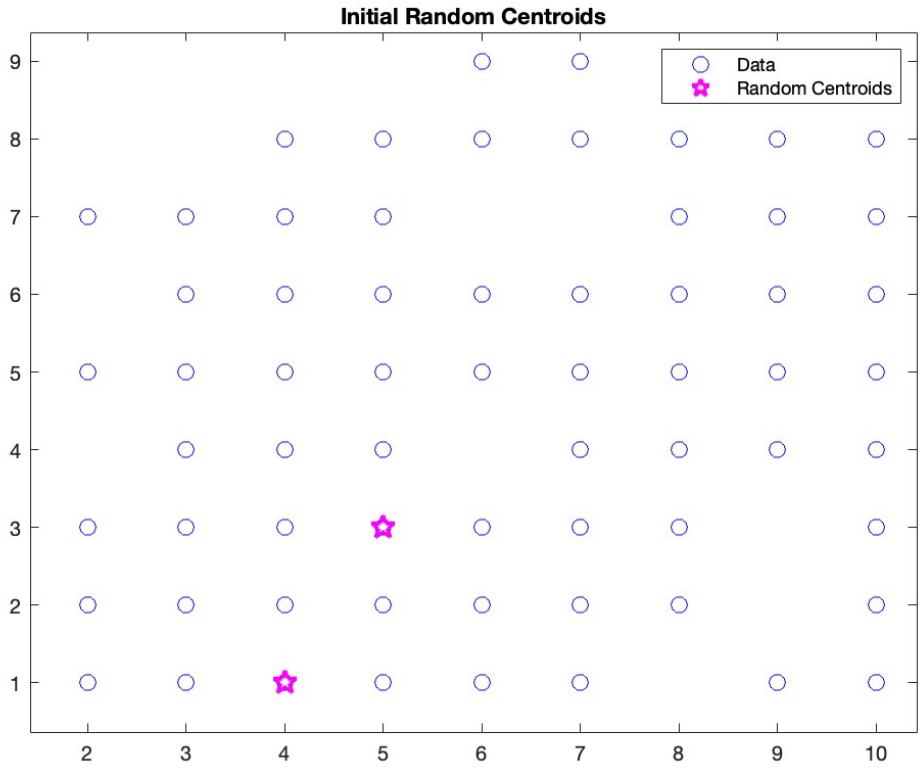

Abbildung 15: Random Medoids

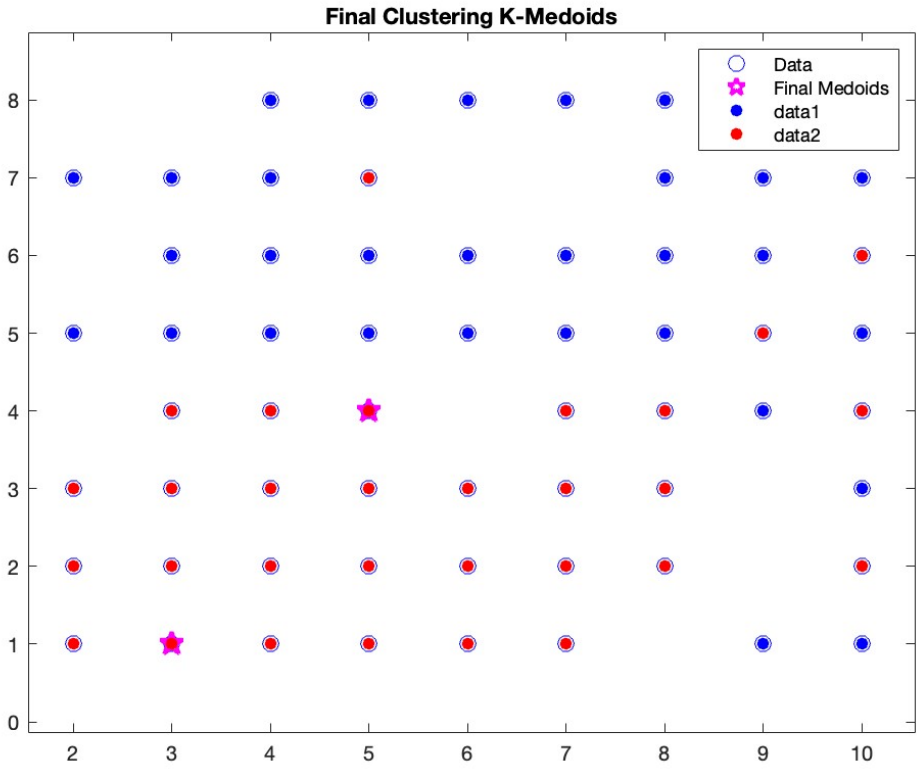

Abbildung 16: Final Medoid Norm 1

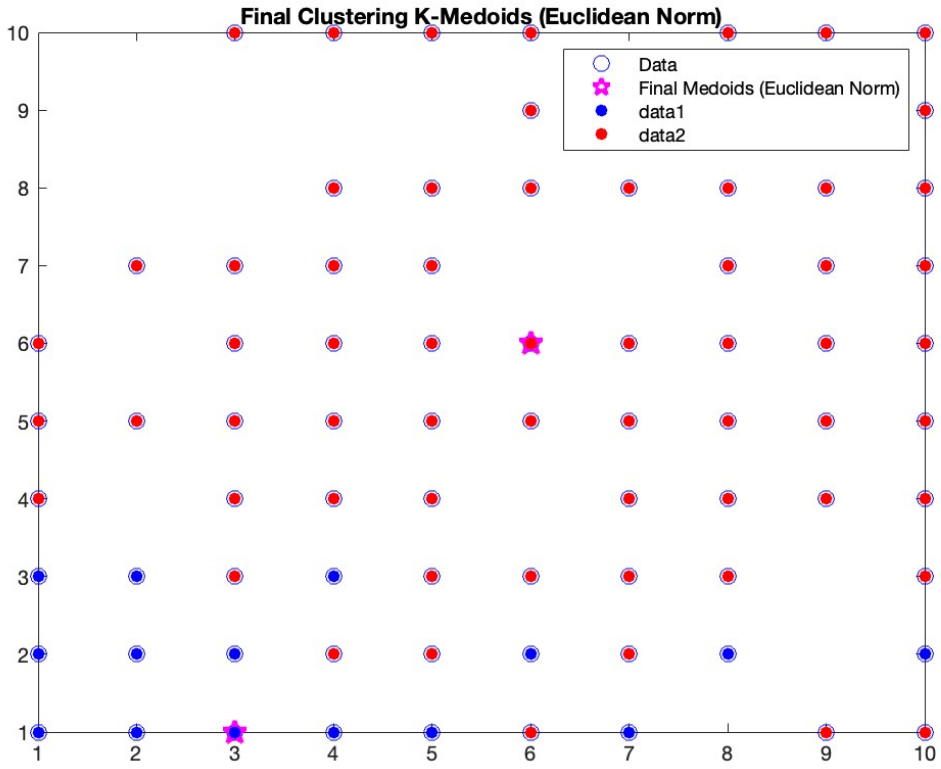

Abbildung 17: Final Medoid Norm 2

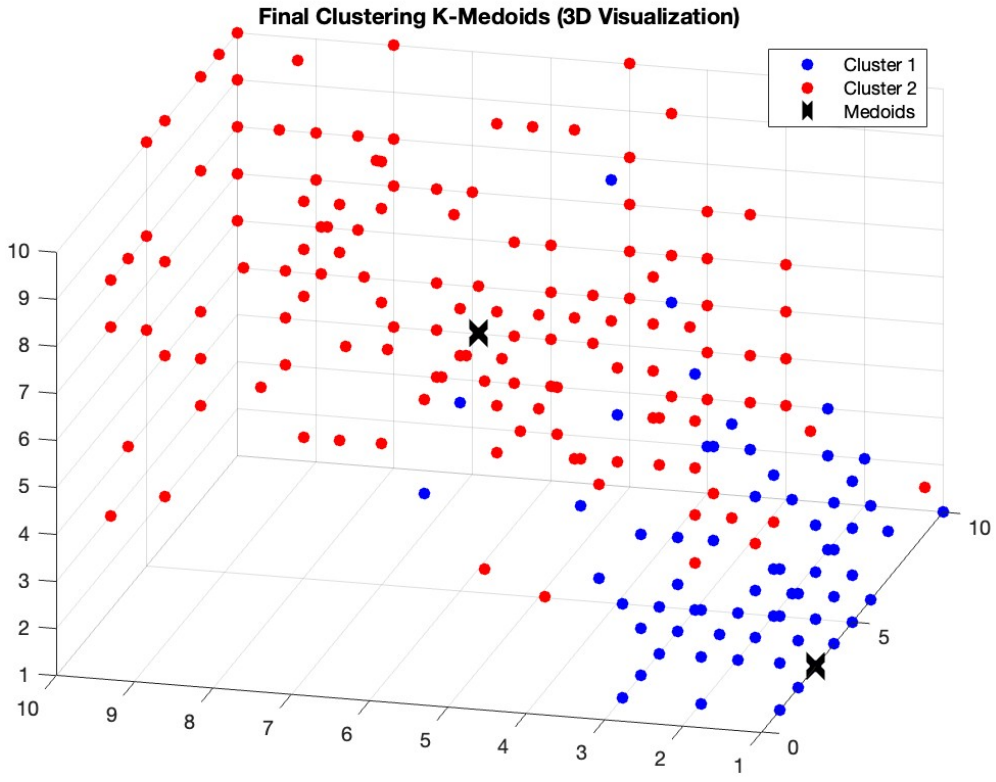

Abbildung 18: Final Cluster Norm 1

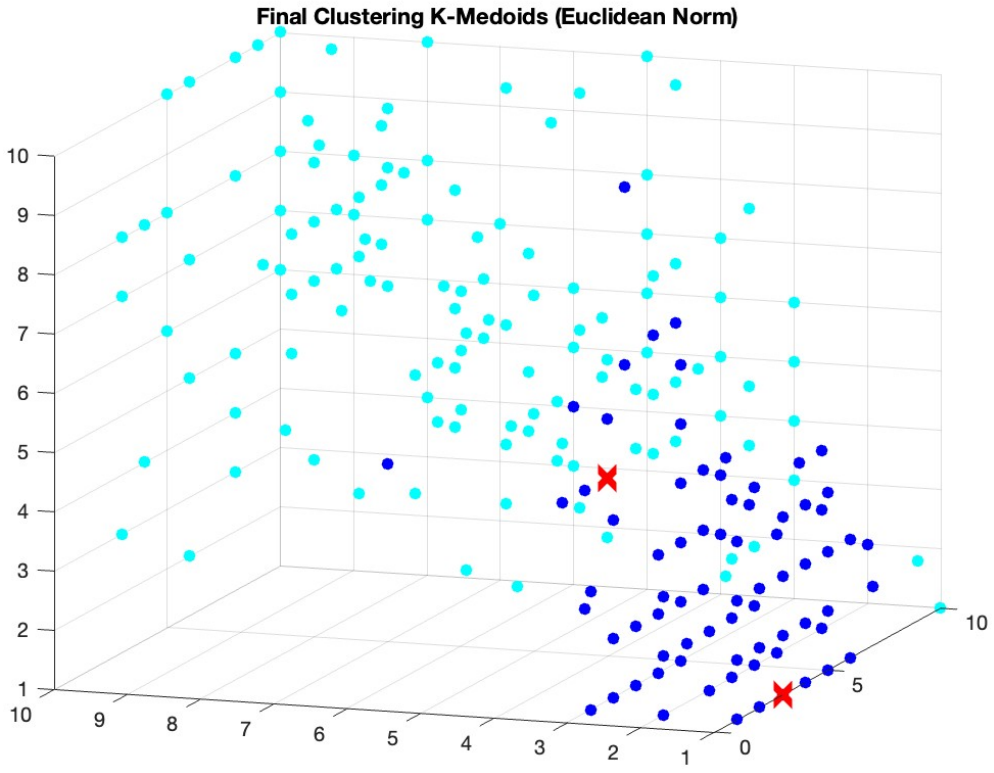

Abbildung 19: Final Cluster Norm 2

3 Question 3(Congressional Vote Dataset)

Download the attached Matlab data file CongressionalVoteData.mat, originating also from the UCI Machine Learning Repository, In your workspace, you find a data matrix X of size 16 × 435, containing the votes of the 435 congressional representatives in 1984 on 16 issues, as well as a vector I of length 435 indicating their partisan membership of each representative (Republican = 0, Democrat= 1).

The columns of the Matrix X give a yes/no vote of each congressional representative on the following 16 issues (“yes” = 1, “no” = 1, “missing vote” = 0):

(a) Write the distance matrix between the representatives, using a dissimilarity index between the “yes” and “no” votes as a distance measure. To address the missing votes, you can follow the guidelines below: – The data contain one representative who did not vote one single time. Discard that representative. – When computing the dissimilarity index between two representatives, include only the issues on which both of them voted. Hence, the dissimilarity index is the number of times the votes disagreed divided by all votes that both candidates cast. – You will find some pairs of representatives with no simultaneous voting record: one or the other was always absent. Use a neutral value 1/2 as a dissimilarity measure for those pairs.

3.1 Answer:

The k-medoids algorithm with both Manhattan and Euclidean distances produced clustering results with significant misclassification rates. The sensitivity and specificity were relatively low, indicating that the algorithm struggled to distinguish between Republicans and Democrats effectively based on their voting records. This suggests that the voting patterns may not be distinct enough for the k-medoids algorithm to separate the two parties accurately using the given data.

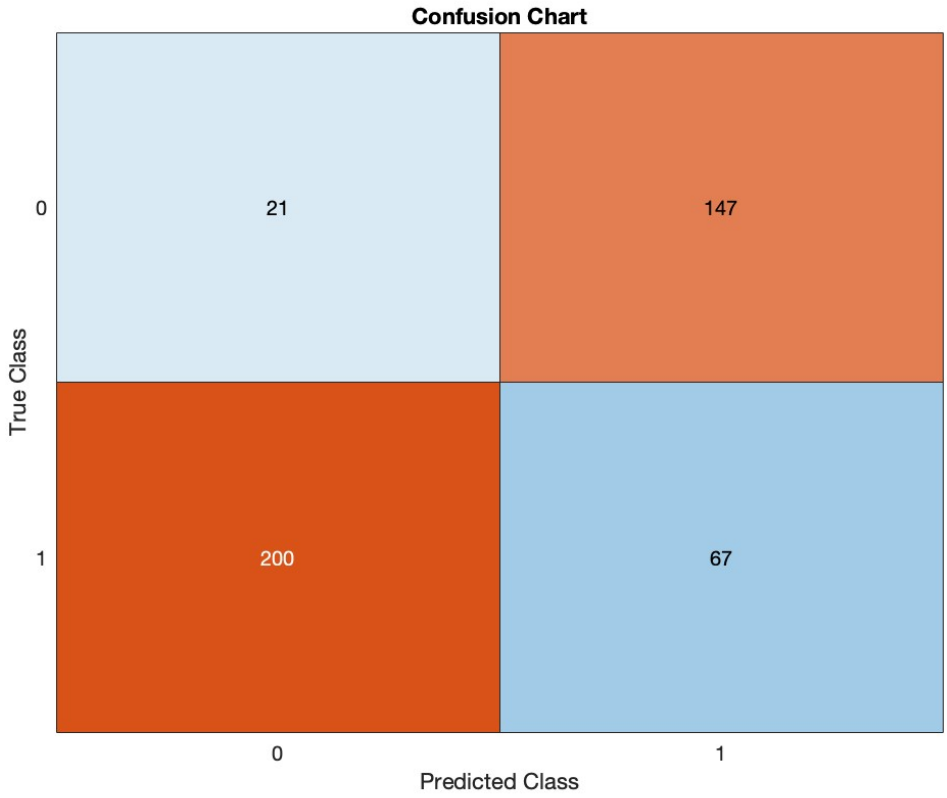

Abbildung 20: Confusion Chart For CongressVote

4 Question 4(Wine Dataset)

Test your k-means and k-medoids algorithm with k = 3 on the data file WineData.mat containing a data matrix X of 13 rows and 178 columns of chemical analysis data of wines derived from the same area of Italy but originating from three different culti- vars. The attributes, corresponding to the rows of the data matrix X are concentra- tions/levels of the following substances:

1 Alcohol

2 Malic acid

3 Ash

4 Alcalinity of ash

5 Magnesium

6 Total phenols

7 Flavonoids

8 Nonflavonoid phenols

9 Proanthocyanins

10 Color intensity

11 Hue

12 OD280/OD315 of diluted wines

13 Proline

Comments on how well, or badly, your algorithms were able to cluster the data to correspond the three different cultivars by comparing your results with the true an- notation recorded in the vector I included in the data file WineData.mat. Report whether the three wine types were easy to cluster based on the recorded attributes.

4.1 Answer:

In exploring the Wine dataset, both the k-means and k-medoids methods showed okay clustering results with k=3, but they weren't perfect. They got some things right, but not everything. Principal Component Analysis (PCA) didn't do well either; it didn't capture the data's structure effectively

We chose Linear Discriminant Analysis (LDA) because it's adept at maximizing the differences between classes while minimizing variations within each class. In essence, it helps separate the different types of wine based on the attributes we provided.

When we applied LDA to the Wine dataset, it performed admirably. By maximizing differences between wine types and minimizing variations within each type, LDA achieved a flawless classification, correctly assigning all wines to their respective types with 100% accuracy. This significantly improved our clustering results compared to other methods like k-means, k-medoids, or PCA.

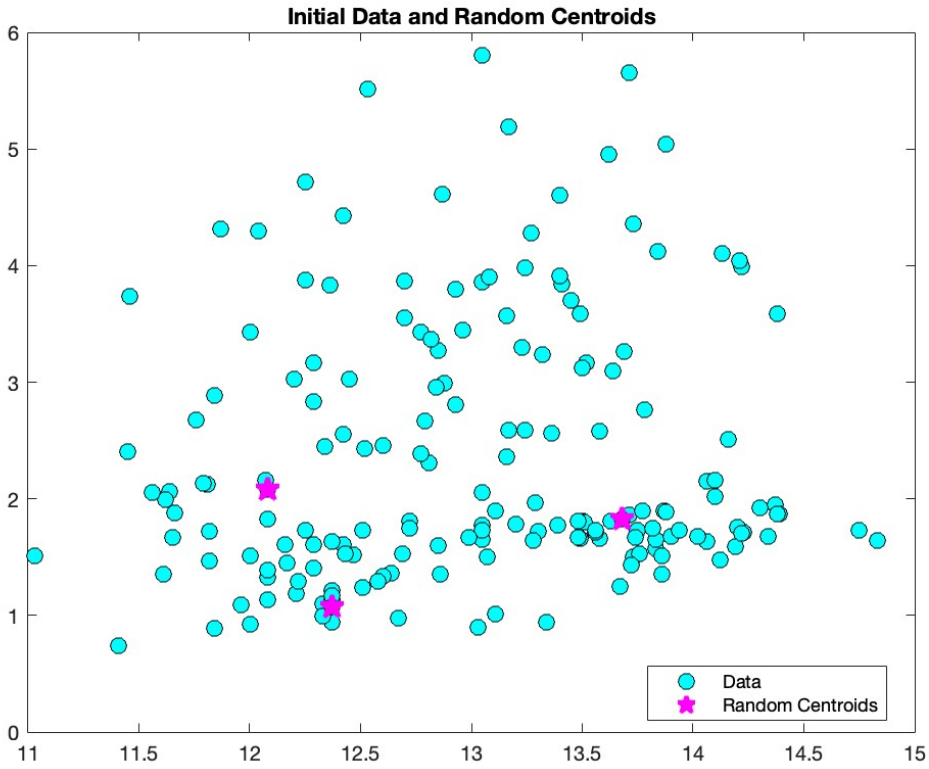

Abbildung 21: Random Medoids

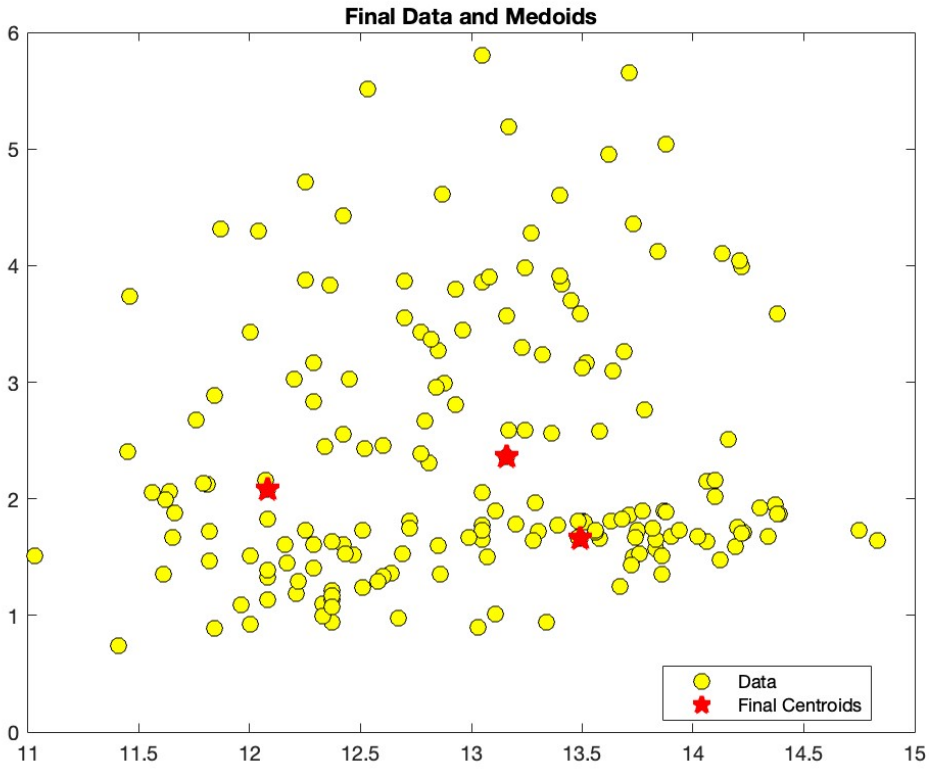

Abbildung 22: Final Medoids

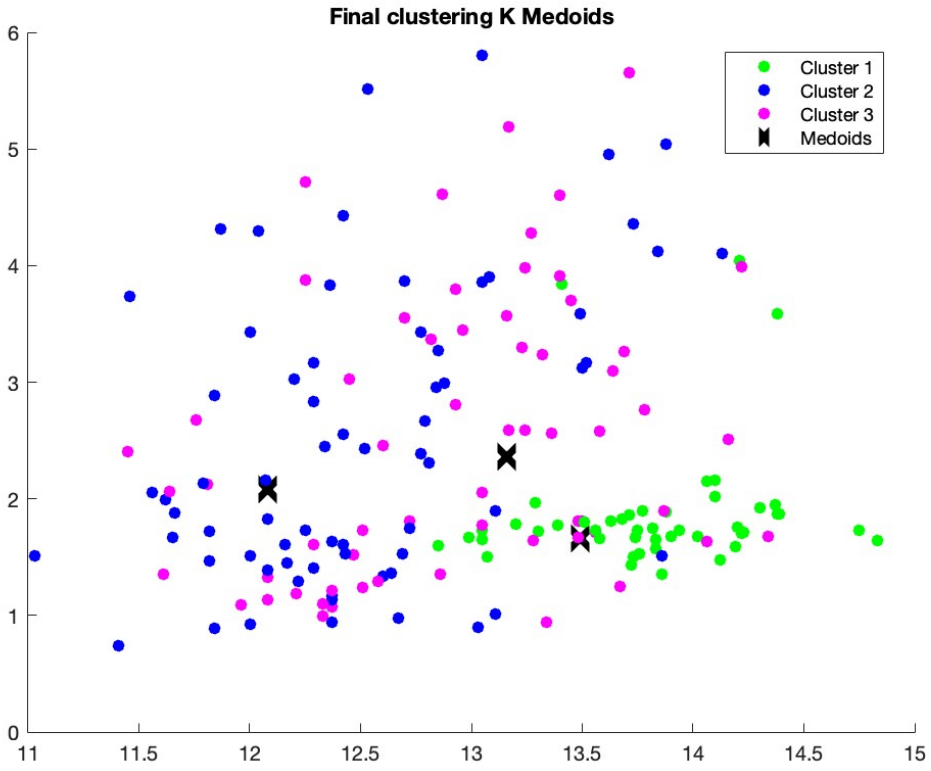

Abbildung 23: Final Clusters and Medoids

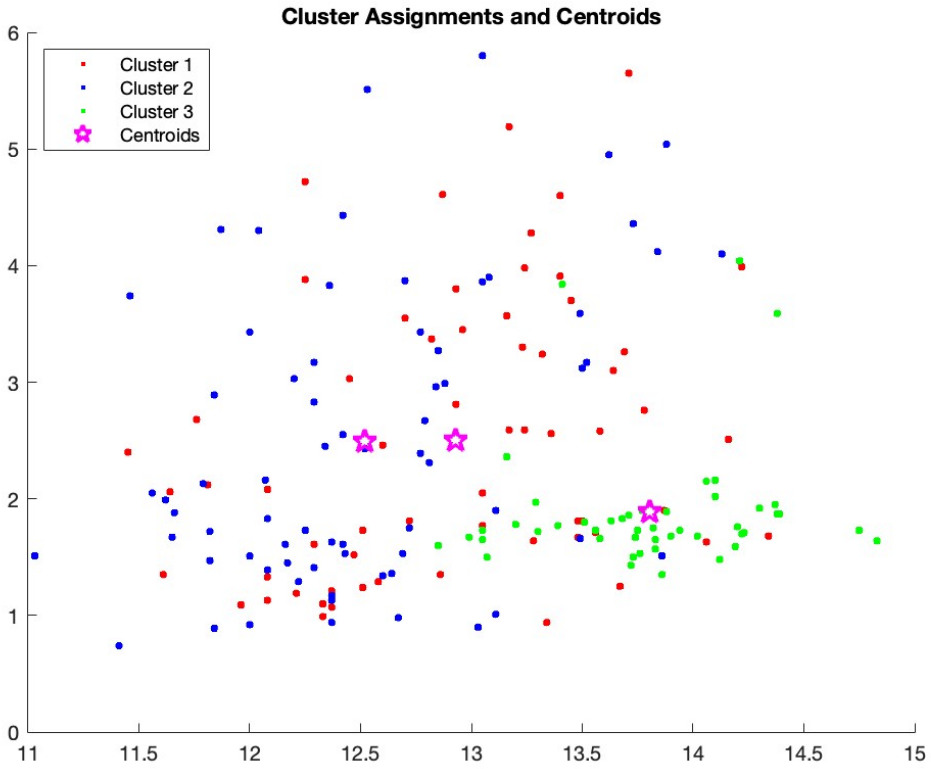

Abbildung 24: Centroids K-mean

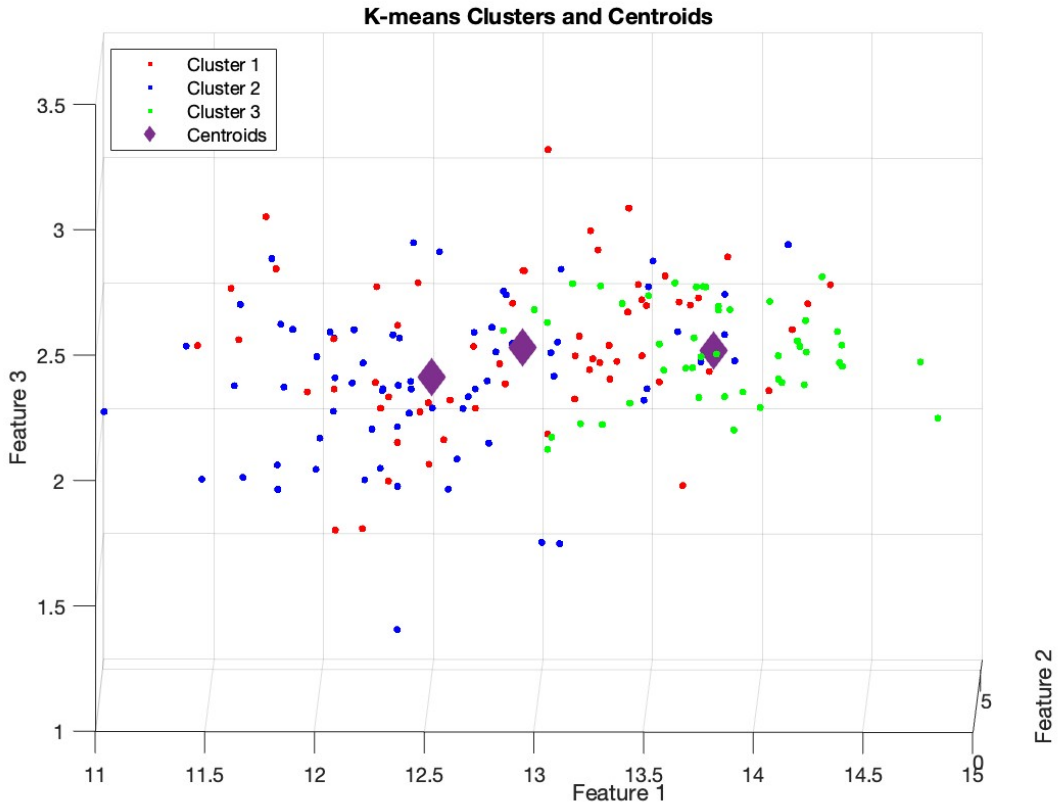

Abbildung 25: Final Clusters and Centroids K-mean

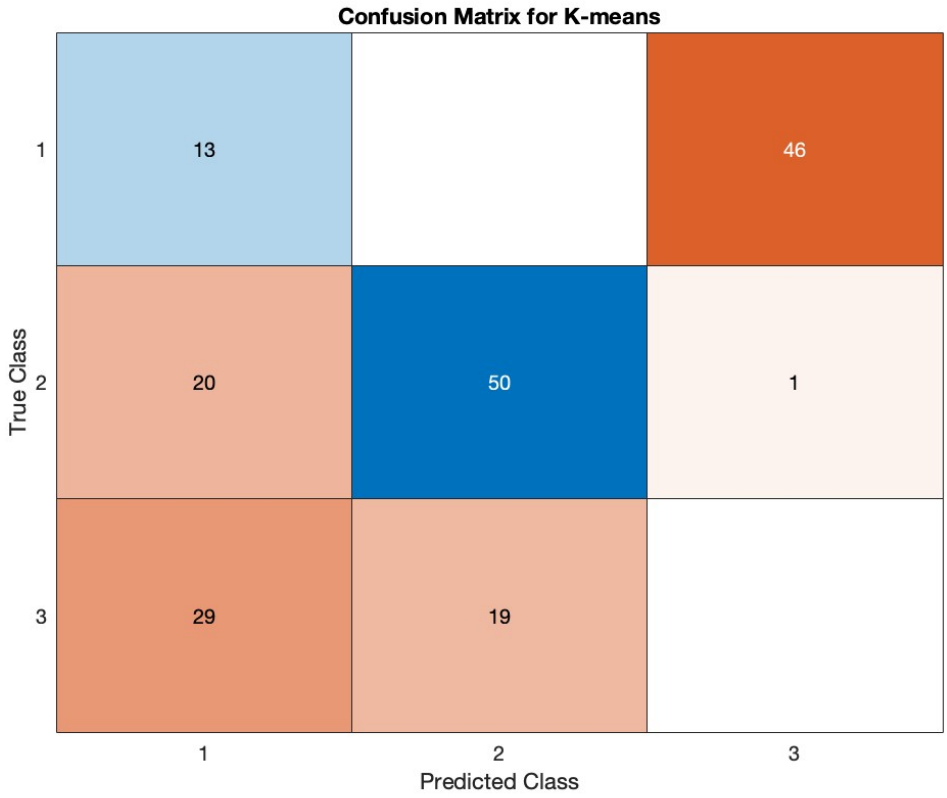

Abbildung 26: Confusion Chart K-mean

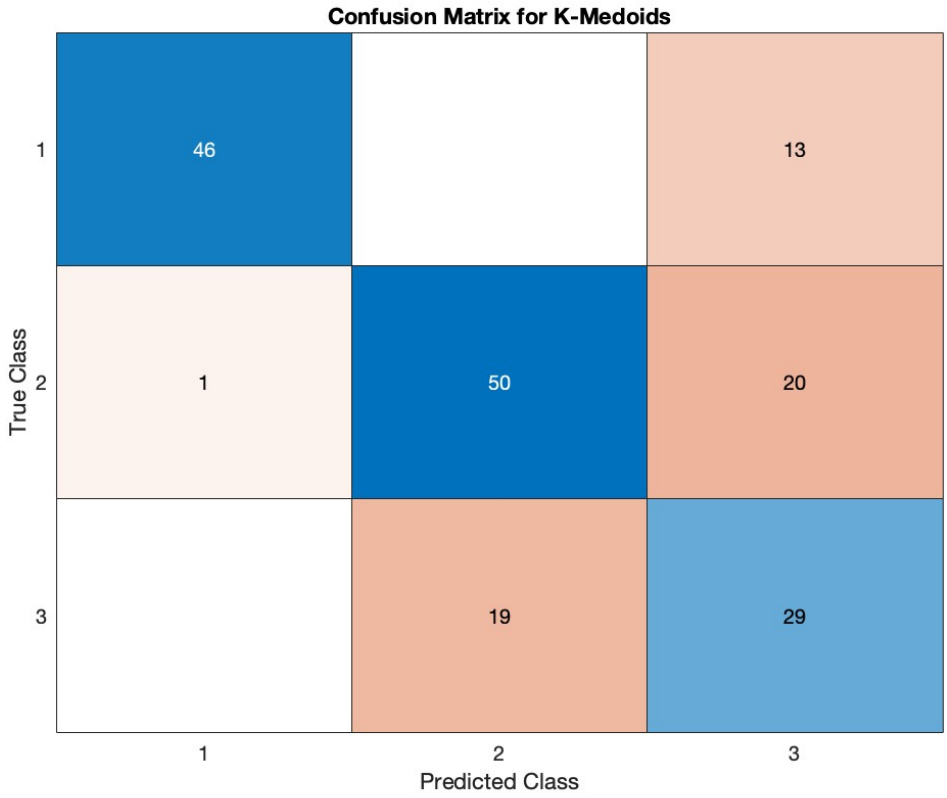

Abbildung 27: Confusion Chart K-medoid

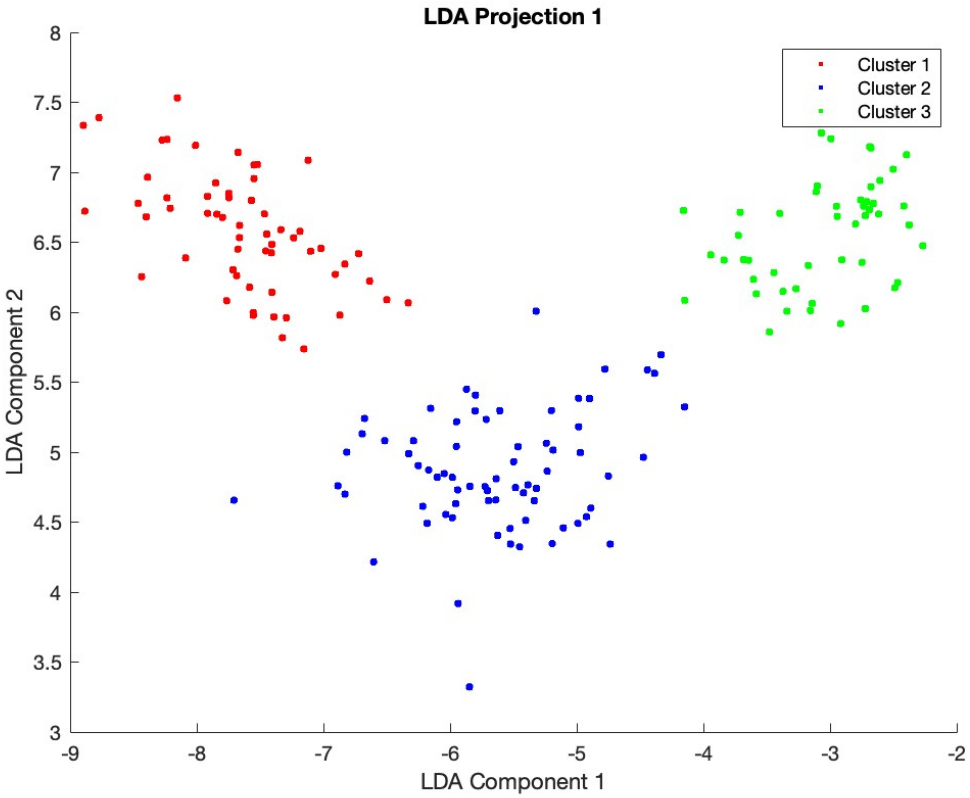

Abbildung 28: LDA Projection First and Second Components

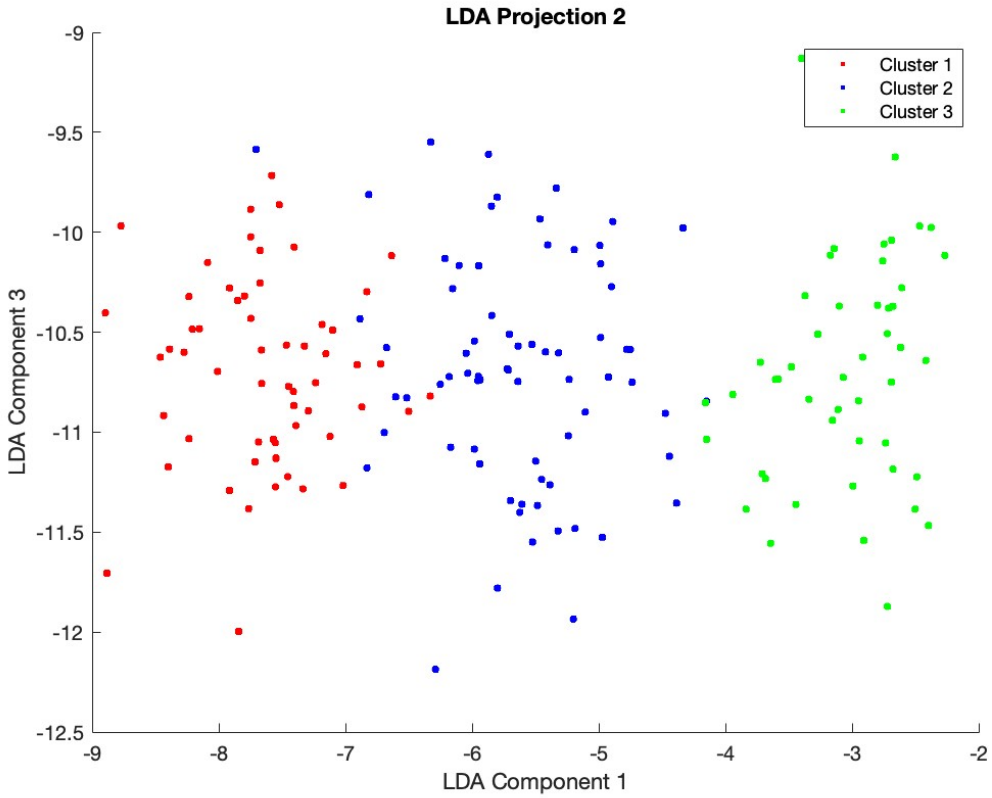

Abbildung 29: LDA Projection First and Third Components

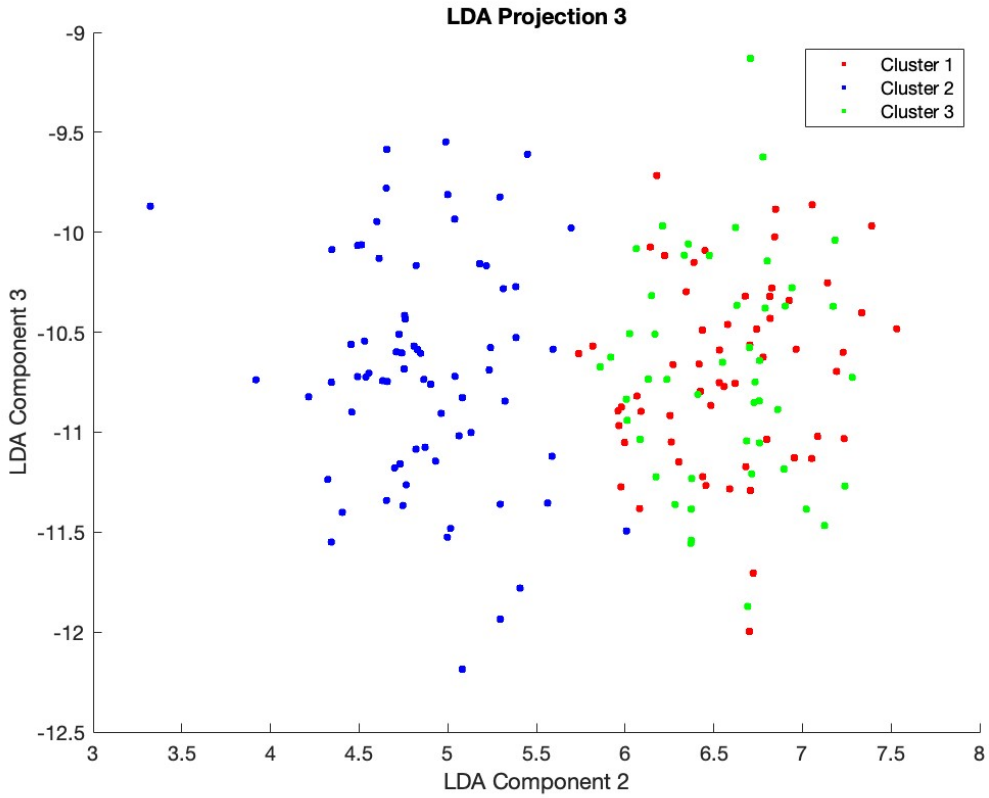

Abbildung 30: LDA Projection Second and Third Components

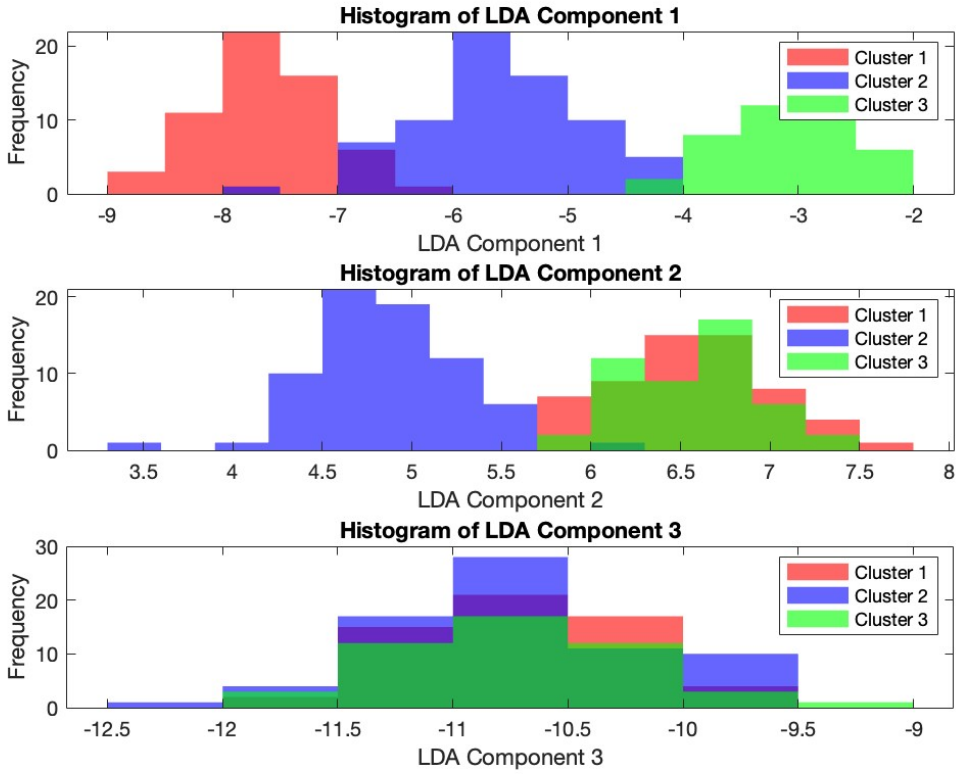

Abbildung 31: Histogram of LDA Components

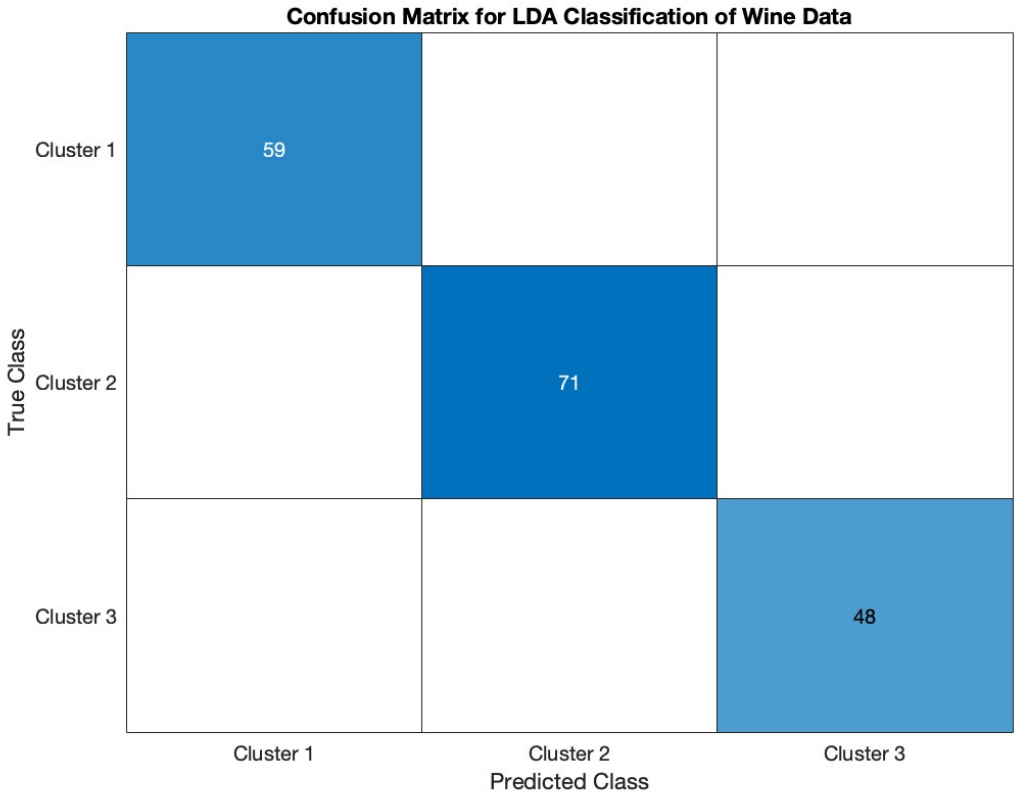

Abbildung 32: Confusion Chart for LDA

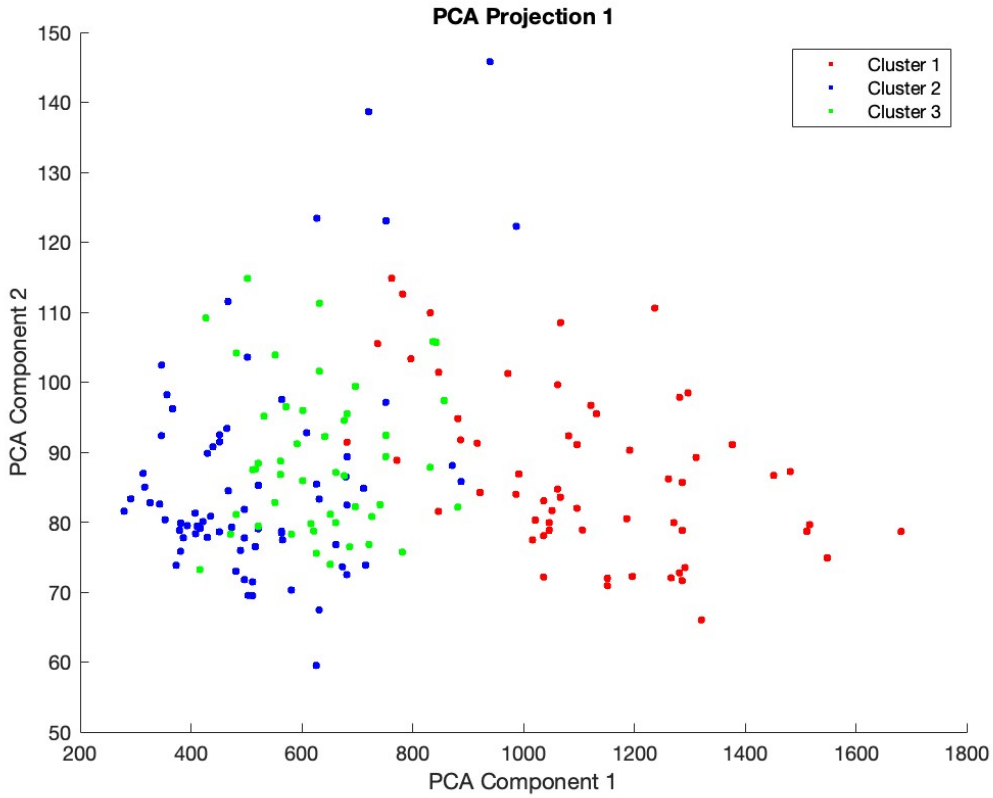

Abbildung 33: PCA Projection First and Second Components

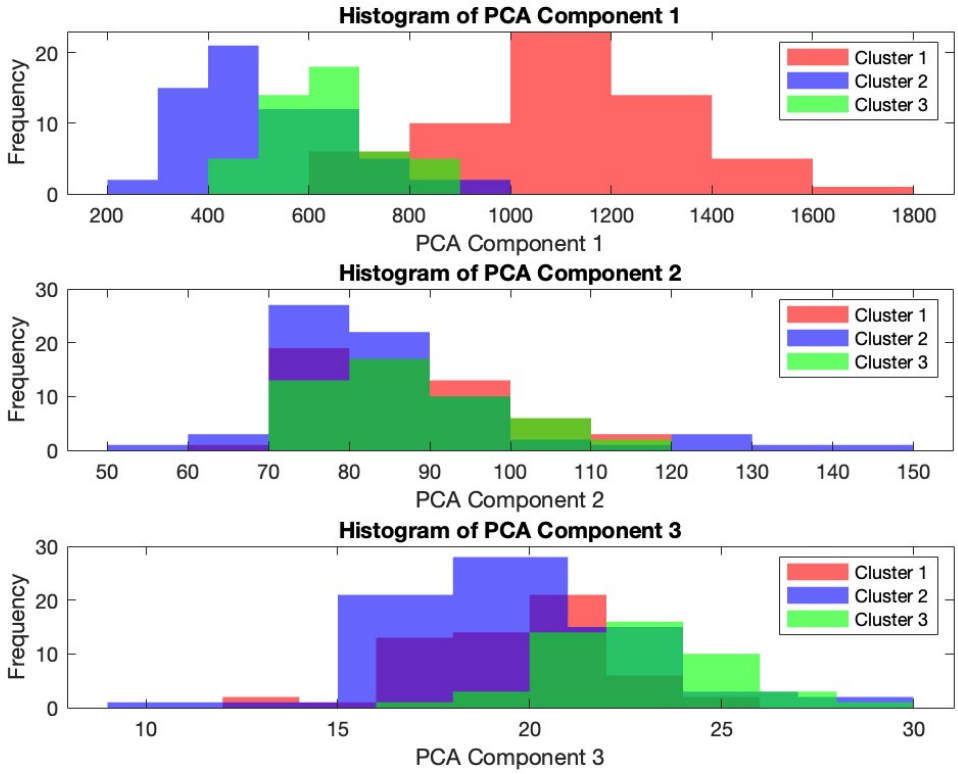

Abbildung 34: Histogram of PCA Components

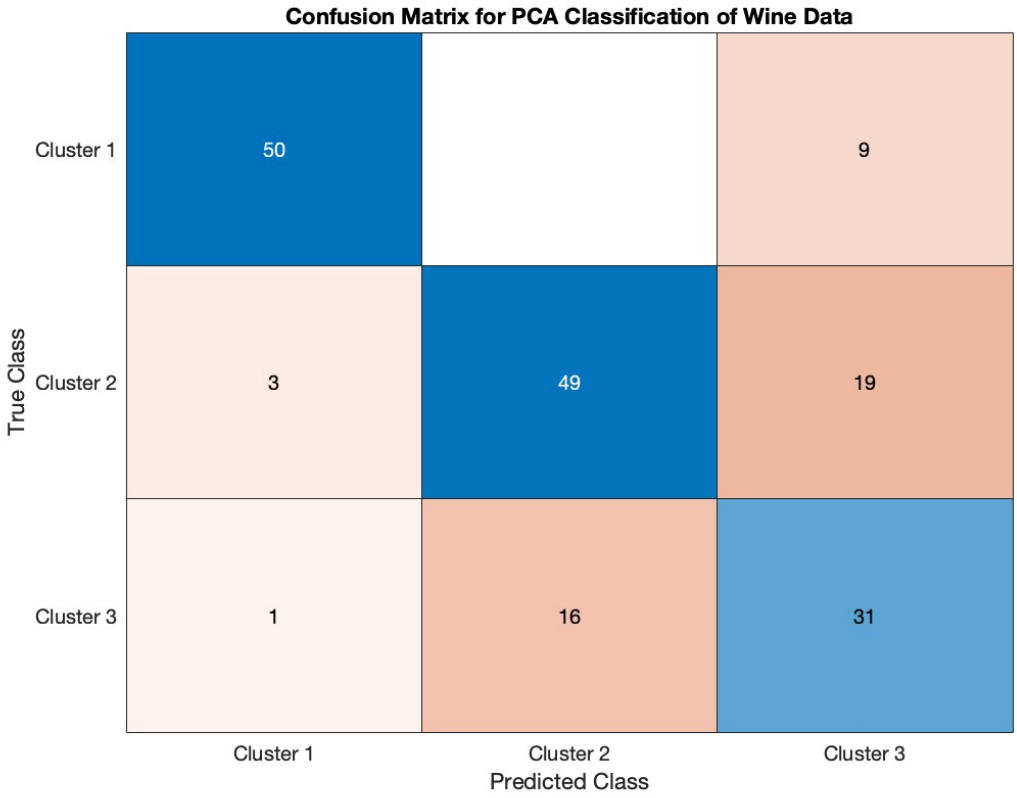

Abbildung 35: Confusion Chart for PCA

5 Question 5(Cardiac Dataset)

Download the data file CardiacSPECT.mat, originating from the UCI Machine Learning, The file contains a binary matrix X of size 22 × 187 and a vector I of length 187. The dataset reports 22 attributes extracted from cardiac Single Proton Emission Computed Tomography (SPECT) images of 187 patients. Each patient was classified into one of two categories: normal or abnormal.

The database of 187 SPECT image sets (patients) was processed to extract features that summarize the original SPECT im- ages.

As a result, 44 continuous feature pattern was created for each patient. The pattern was further processed to obtain 22 binary feature patterns.

The vector I, also in binary form, contains the annotation of the corresponding patient in one of the groups.

(*a*) Write the distance matrix between the patients, using a dissimilarity index between the 0 and 1 as a distance measure. (*b*) Once you have the distance matrix, run the k-medoids algorithm to cluster the patients in two groups, A and B. To investigate how well - or badly - your clustering corresponds to the classification given by the cardiologist, write a matrix C of size 2 × 2 with the following entries:

c11 = number of 1's in your cluster A,

c12 = number of 1's in your cluster B,

c21 = number of 0's in your cluster A

c22 = number of 0's in your cluster B.

Comment on how the matrix C still provides you some insight on the degree of agreement between the clustering that you found and the classification by the cardiologist if

c11 = number of 1's in your cluster B,

c12 = number of 1's in your cluster A,

c21 = number of 0's in your cluster B

c22 = number of 0's in your cluster A.

In the light of your clustering, how well do the attributes represent the state of the patients?

5.1 Answer:

The final medoids identified by the k-medoids algorithm are patients 20 and 44. These serve as the central patients for the two clusters formed.

$c11$=38: Number of abnormal patients in cluster A. $c12$=46: Number of abnormal patients in cluster B. $c21$=68: Number of normal patients in cluster A. $c22$=35: Number of normal patients in cluster B.

The k-medoids algorithm identified patients 20 and 44 as the central patients for the two clusters formed. This clustering resulted in a balanced distribution of abnormal and normal patients within each cluster, showing some alignment with the cardiologist's classifications. However, significant discrepancies remain, underscoring the challenges of clustering medical data.

The k-means results, with very large numbers, indicate a broad spread of patients across clusters, but these numbers are harder to interpret without normalization.

Given the limitations of k-means and k-medoids, we turned to Linear Discriminant Analysis (LDA), which showed improved results. However, perfection wasn't attained, possibly due to the complexity of working with medical data.

Comparing LDA with Principal Component Analysis (PCA), LDA emerged as the superior method, although not flawless. The dataset's attributes capture crucial aspects of patient states, enabling meaningful clustering. Yet, imperfections imply the need for additional features or refined methods to enhance accuracy.

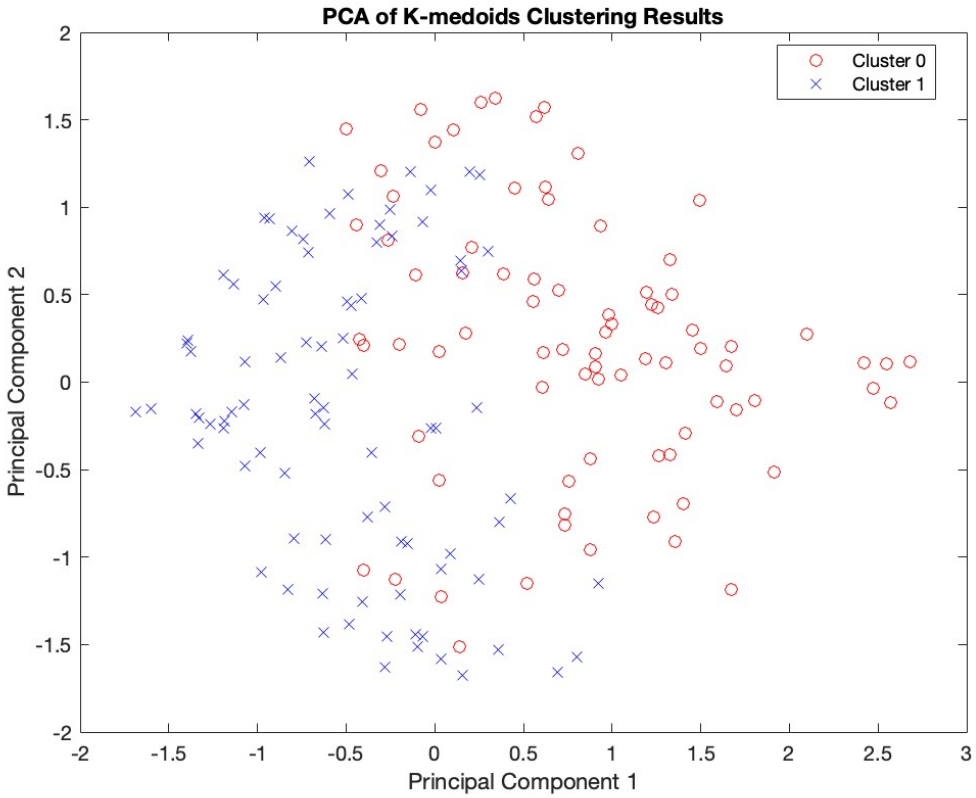

Abbildung 36: Final Clusters K-medoid

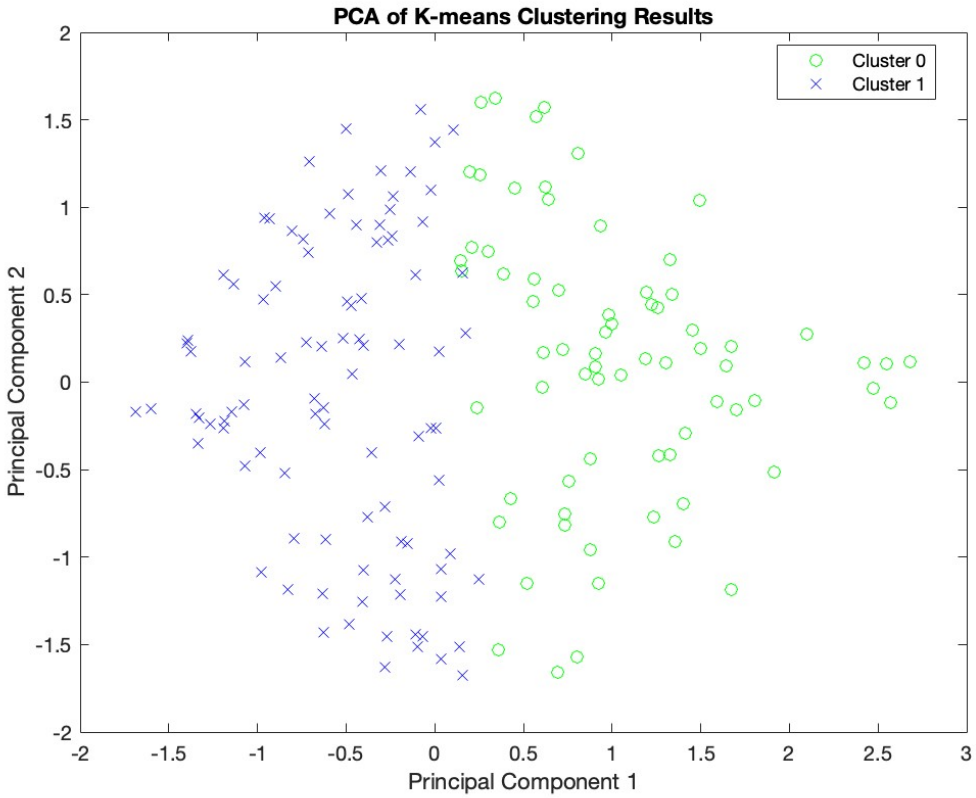

Abbildung 37: Final Clusters K-mean

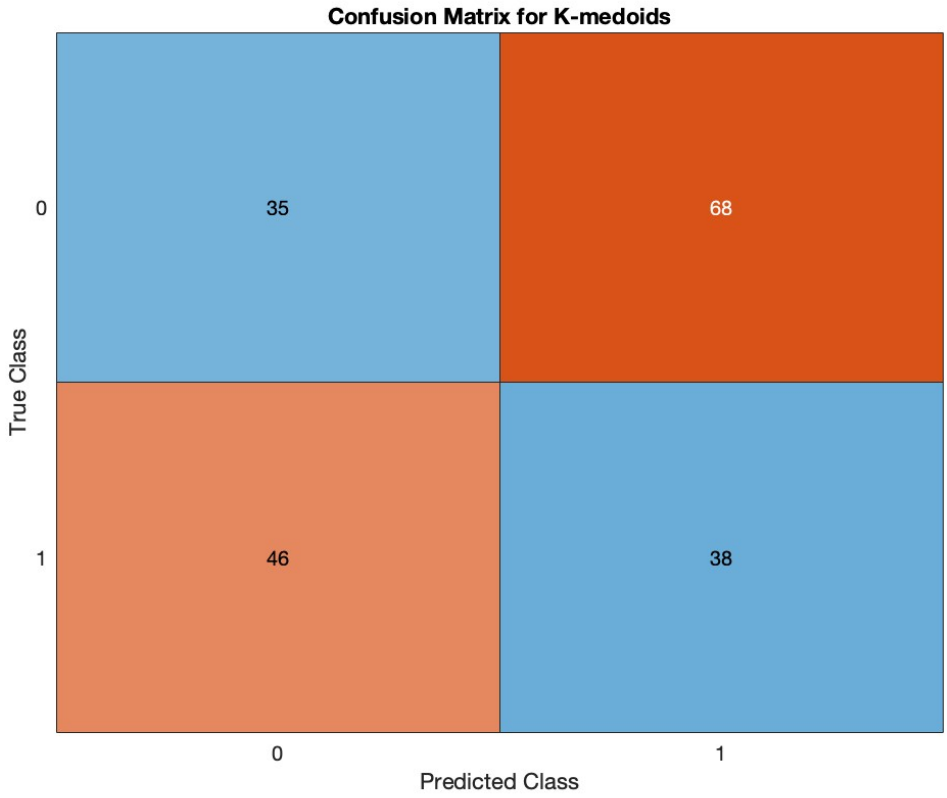

Abbildung 38: Confusion Chart For K-medoid

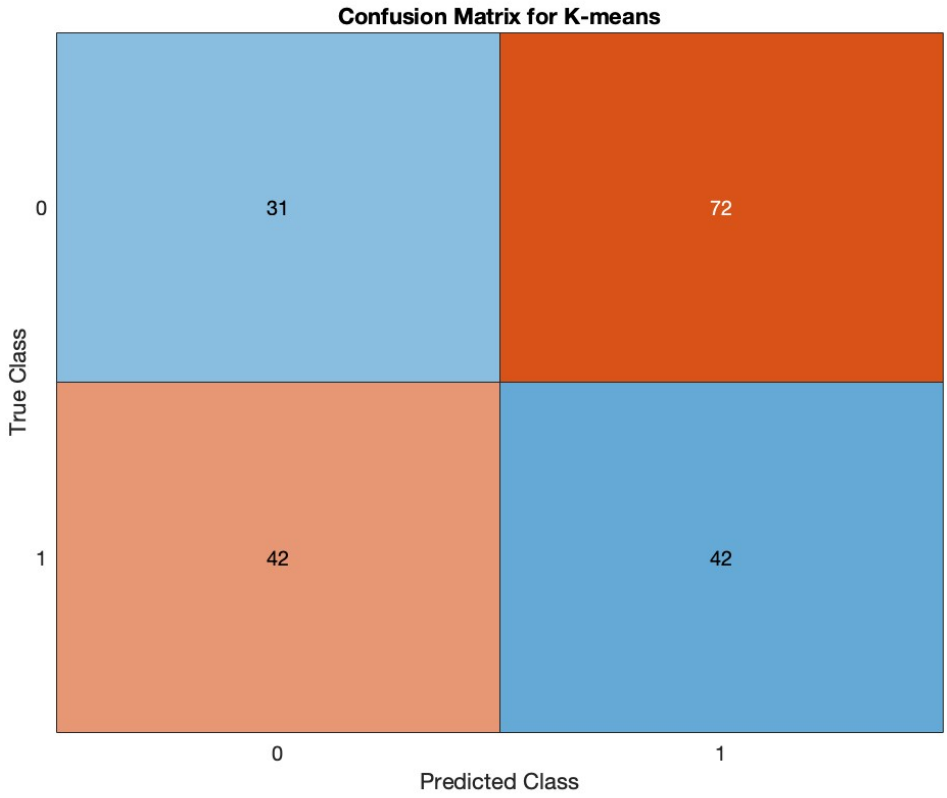

Abbildung 39: Confusion Chart For K-mean

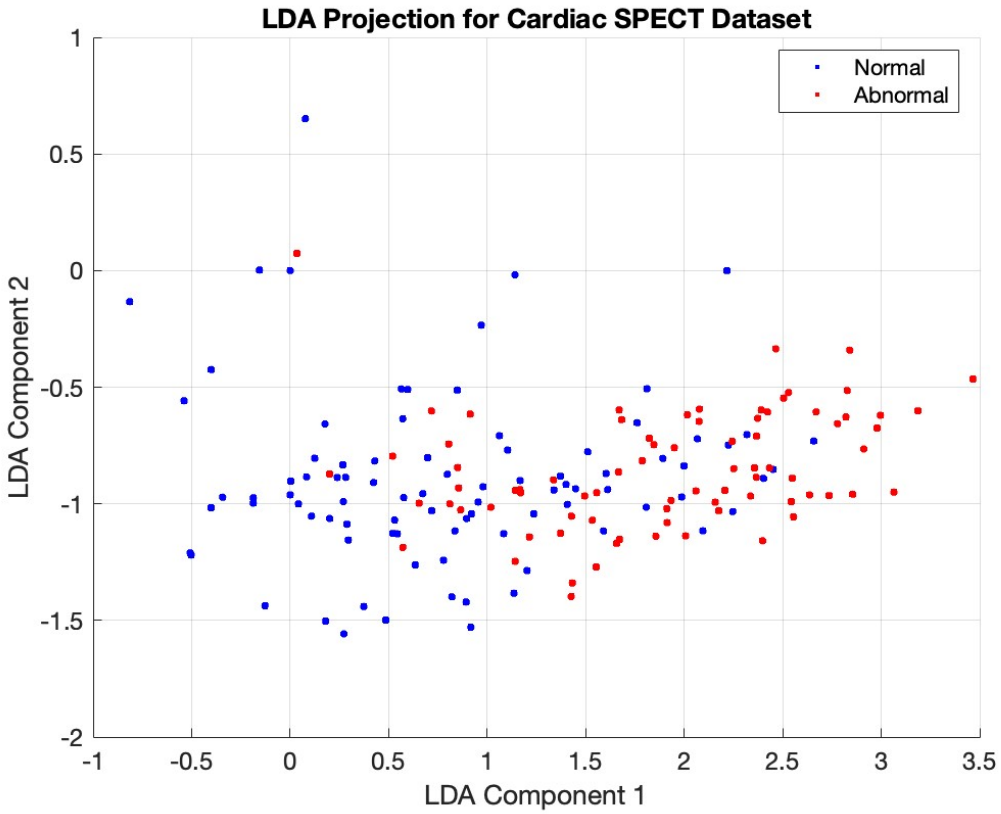

Abbildung 40: LDA Projection

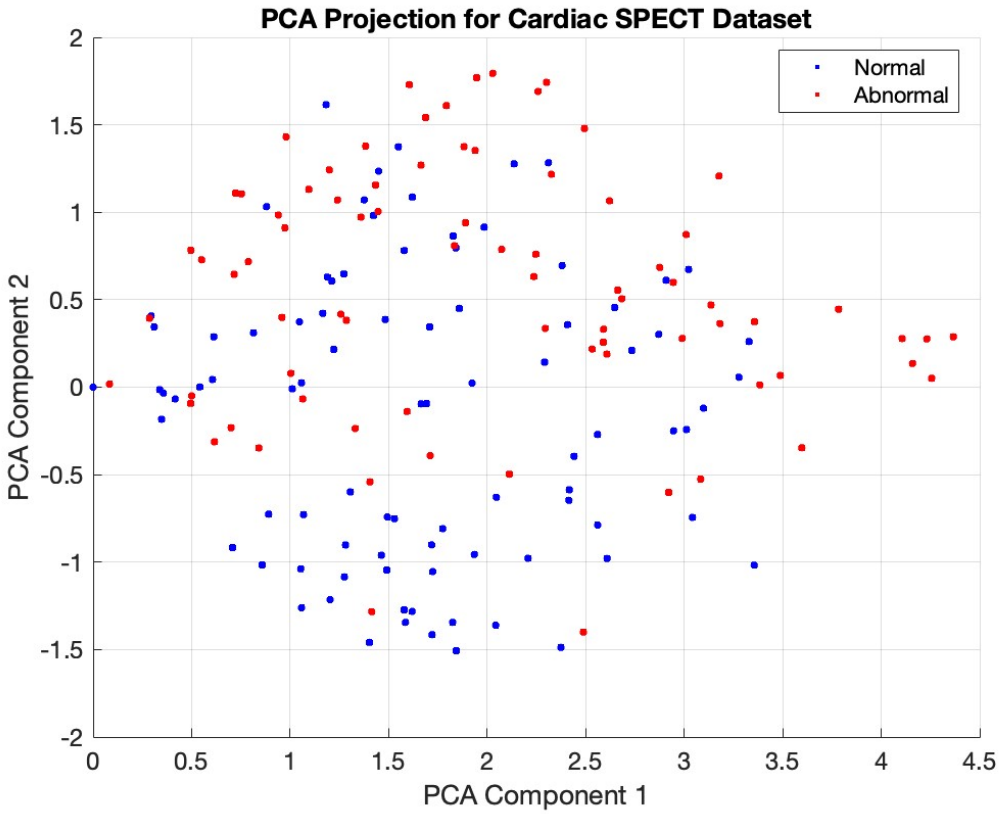

Abbildung 41: PCA Projection

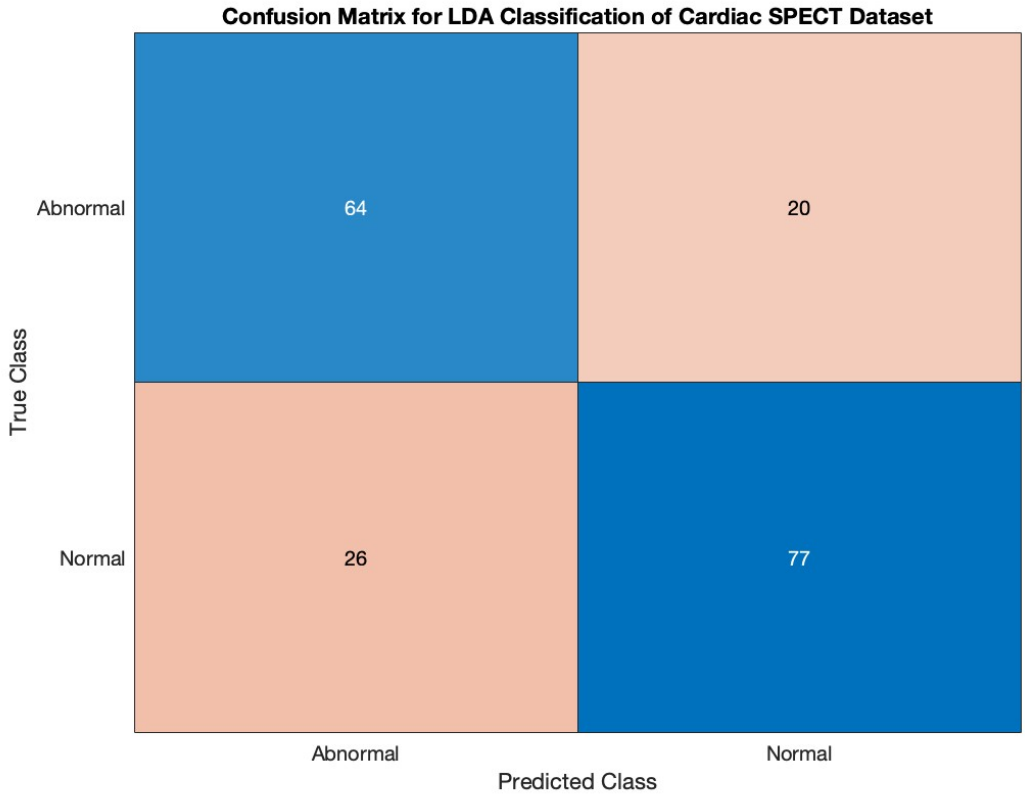

Abbildung 42: Confusion Chart For LDA

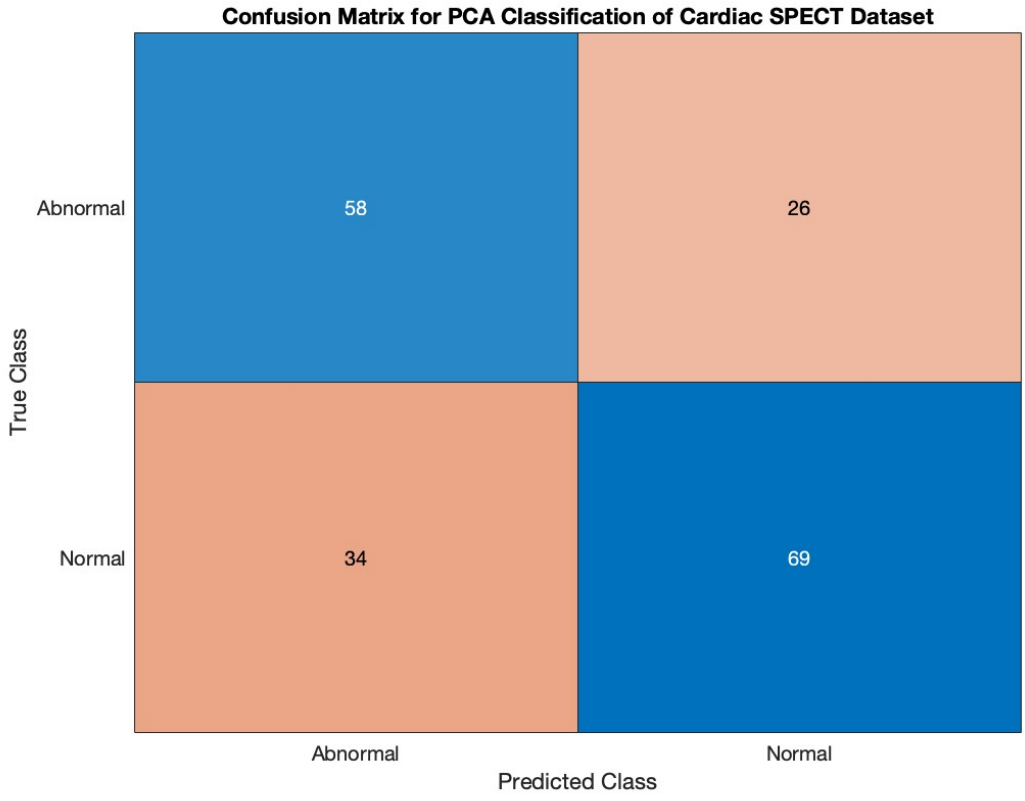

Abbildung 43: Confusion Chart For PCA

6 Question 6 (HandwrittenDigits Dataset)

Using again the handwritten digits data, select two different digits, for instance ’0’ and ’4’. Denote by X(04) the data matrix containing the images.

(*a*) Plot the data projected on the first two LDA directions

(*b*) Write your ANLS algorithm to find the rank-k non-negative factorization.

Run the algorithm with different values of k, e.g., k = 5, 10, 20, and plot the feature

vectors as images. You should see pieces of the digits ’0’ and ’4’ in those images. You may use the built-in Matlab function lsqnonneg in your algorithm. Pick one of the feature vectors that clearly looks like a piece of a ’0’, and verify that the coefficients in the matrix H that correspond to that feature vector reveal if the corresponding data vector represents a ’0’ or a ’4’.

6.1 Answer:

Throughout our analysis, Linear Discriminant Analysis (LDA) emerged as a robust classifier for the given dataset. It effectively distinguished between the digits ’0’ and ’4’, showcasing its discriminative power in separating distinct classes.

By projecting the data onto the first two LDA directions, we observed clear clusters representing the different digits, highlighting the ability of LDA to capture the underlying structure of the data and facilitate accurate classification. The precision of LDA in correctly categorizing the handwritten digits underscores its efficacy as a classifier for this dataset.

Moreover, throughout the Non-negative Matrix Factorization (NMF) process, we noted consistent convergence trends across various k values. Typically, convergence was reached within 20-30 iterations. However, for larger k values such as 20, we observed a slower convergence rate, with a slow decrease in relative changes after approximately 45 iterations. Interestingly, the decrease in relative changes was more pronounced for k=10 compared to k=5, suggesting a nonlinear relationship between k and convergence rate.

In summary, our investigation underscored the efficacy of LDA in distinguishing between the digits ’0’ and ’4’. Furthermore, the successful implementation of the ANLS algorithm for non-negative factorization provided valuable insights into the structure of the data.

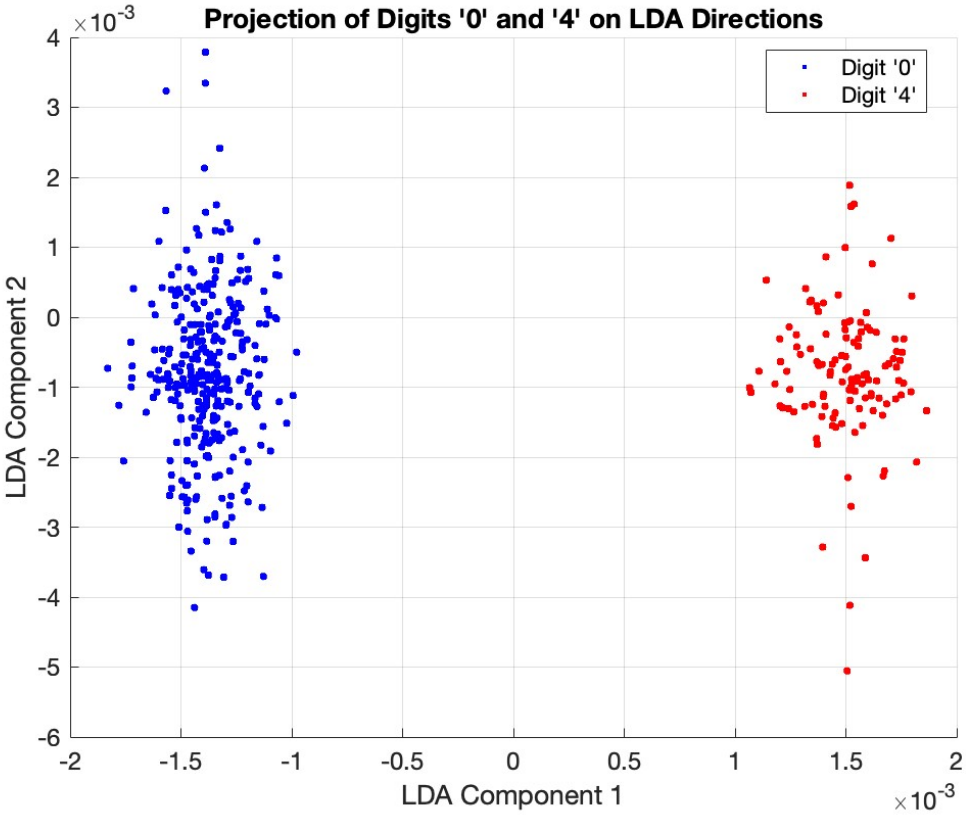

Abbildung 44: LDA Projection

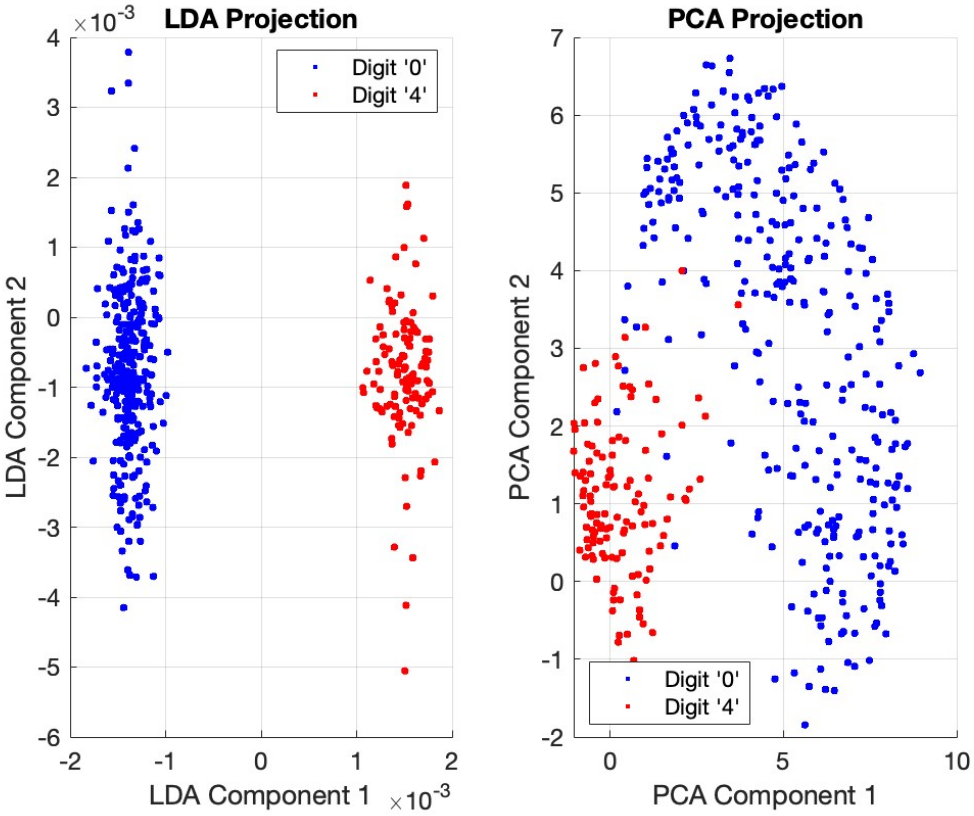

Abbildung 45: LDA vs PCA Projection

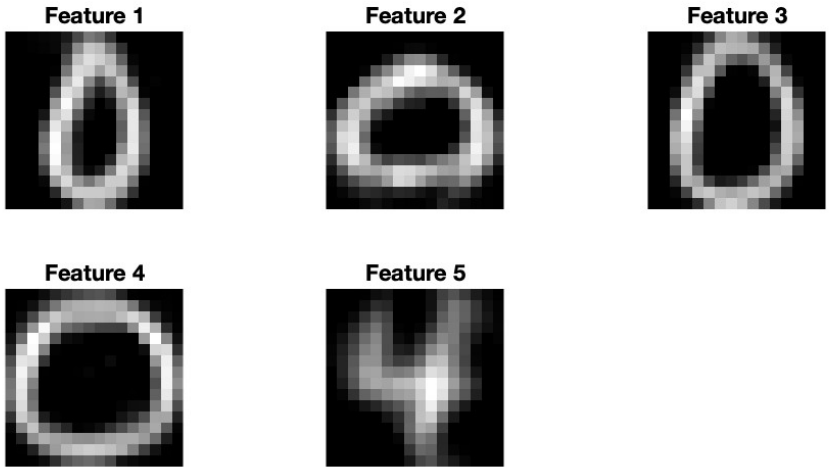

Abbildung 46: Feature Vector k=5

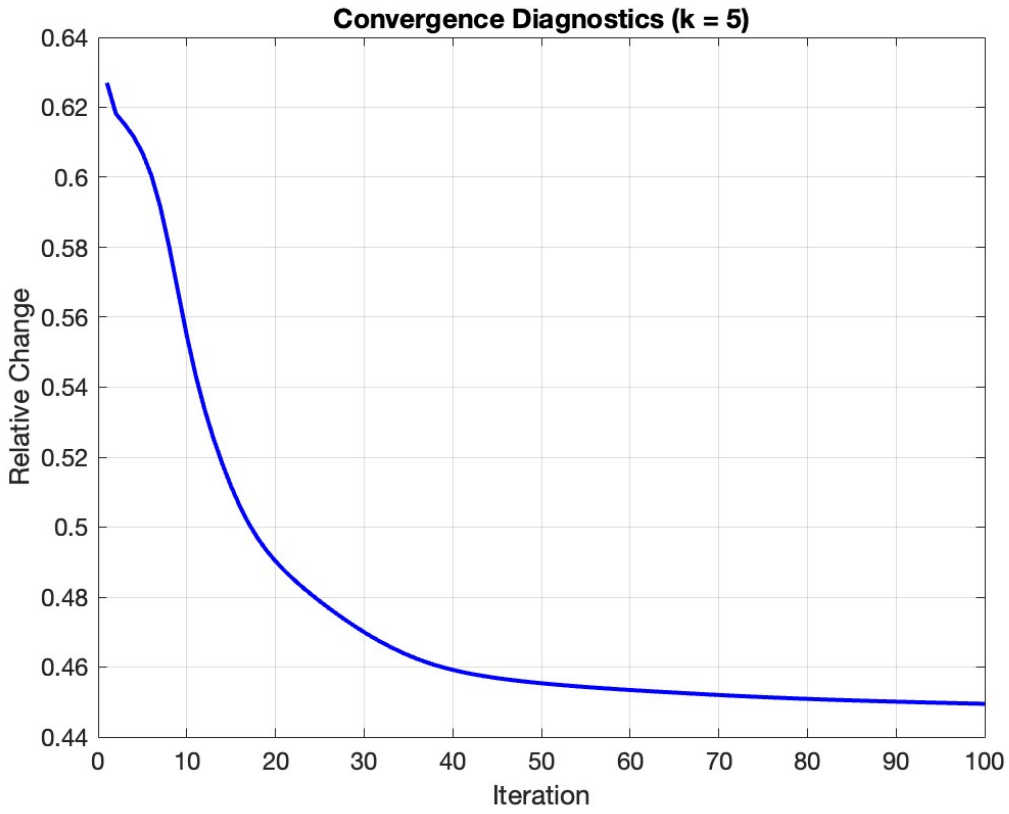

Abbildung 47: Convergence Diagnostic k=5

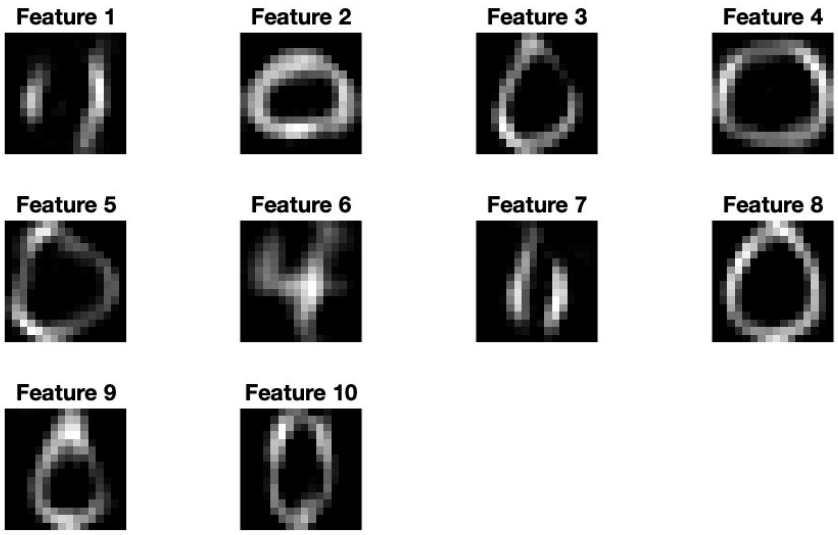

Abbildung 48: Feature Vector k=10

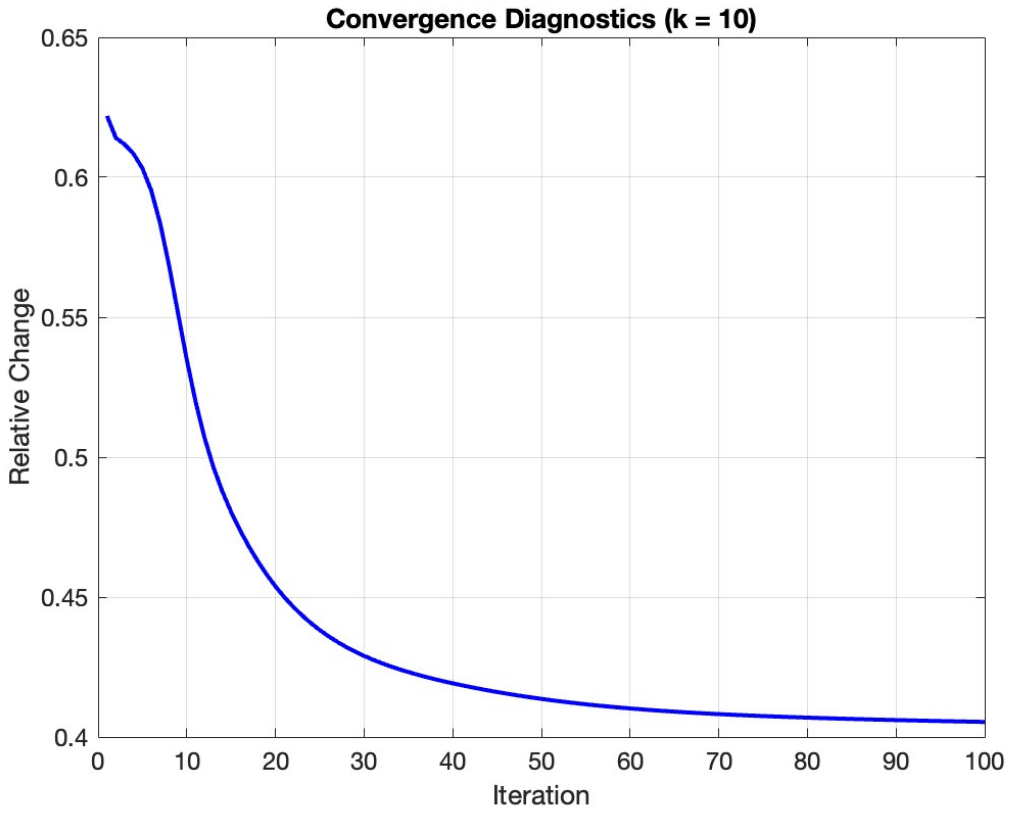

Abbildung 49: Convergence Diagnostic k=10

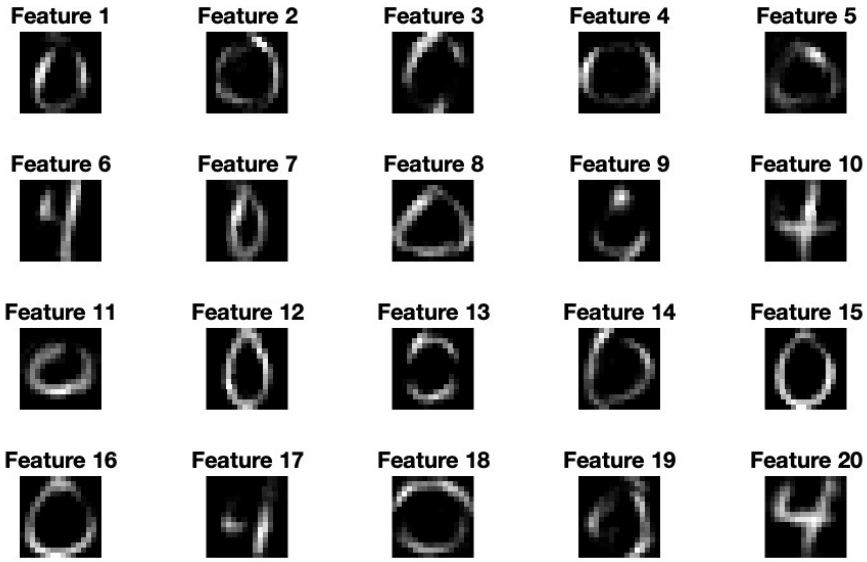

Abbildung 50: Feature Vector k=20

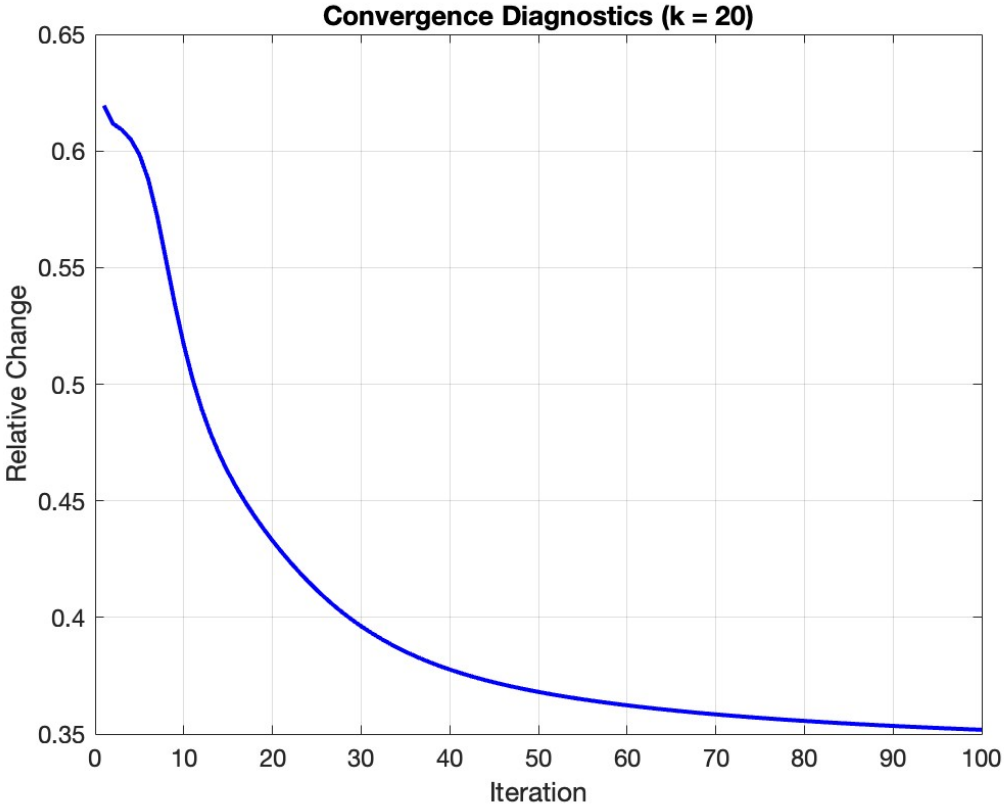

Abbildung 51: Convergence Diagnostic k=20