MPA-MLF - Miniproject

Theme: Classification of wireless transmitters

Date: 31.03.2024

Authors: Marek Černý, Dušan Kratochvíl

Task description

The task of this project is to determinate the number of wireless transmitters based on the hardware imperfections. We were provided with the dataset that consists of 19200 samples and 9 features in total. The features represent the main radio frequency impairments, such as: Carrier Frequency Offset (CFO) between transmitter and receiver, CFO measured after demodulation, gain imbalance of modulator, combination of gain and quadrature imbalance aggregated into one parameter, Origin (DC) offset, quadrature error imbalance, phase difference between received and ideal constellation points, magnitude error between received and ideal constellation points in QAM and error vector magnitude that is representing the RMS error between received and ideal constellation points, magnitude error between received and ideal constellation points in QAM illustrated in Figure 1.

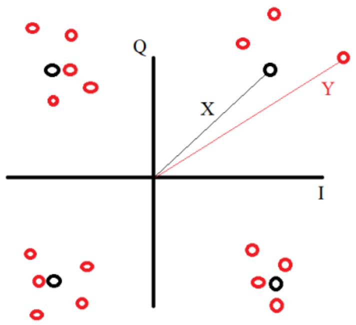

Figure 1: Original (black) and distorted (red) constellation points

The effect of gain imbalance, quadrature error and origin offset on the QPSK constellation is illustrated in Figure 2.

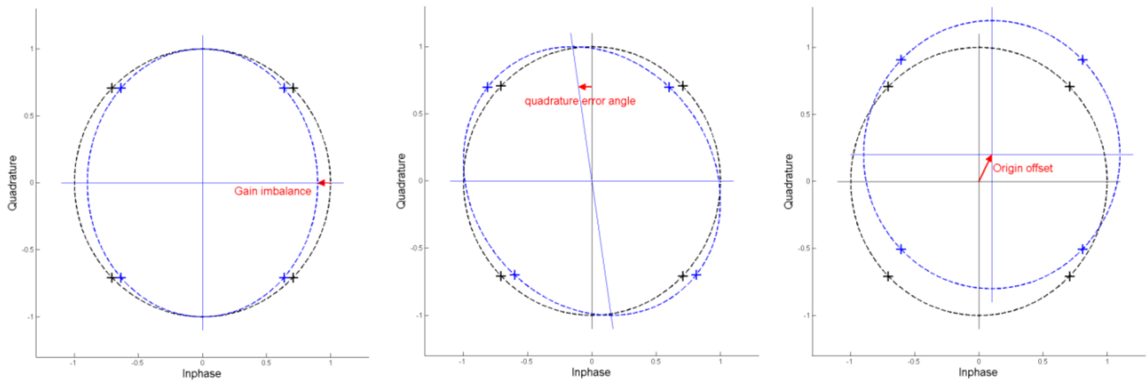

Figure 2: Effect of selected imbalances on QPSK constellation diagram

Solution process

Since transmitters were not part of the dataset, we decided to use the K-means extended by PCA and then display the results using the elbow method.

K-means is one of the most popular machine learning algorithms, well known for its simplicity. This algorithm aims to separate unlabeled data into different clusters with the number of clusters represented by the variable K. The similarity of data is based on some similarity metric that can be distance-based, correlation-based, etc.

One of the disadvantages of using k-means is that it requires a number of centroids k as an input. There are two methods for choosing the right number k, namely elbow method and silhouette analysis.

The Elbow method is a technique used in determining the optimal number of clusters for k-means clustering. The basic idea is to plot the explained variation as a function of the number of clusters and pick the elbow of the curve as the number of clusters to use. The main advantage of this method that its very simple and easy to implement, on the other hand sometimes it is not very clear where the elbow is.

Principal Component Analysis (PCA) is a dimensionality reduction technique used to simplify complex dataset while preserving as much of the variability in the data as possible. It works by transforming the original features of the dataset into a new set of orthogonal features called principal components. The main idea behind PCA is to find directions (principal components) in which the data varies the most. These principal components are computed such that the first principal component explains the maximum variance in the data, the second principal component explains the maximum variance remaining after the first component is accounted for, and so on. Each principal component is a linear combination of the original features.

Cumulative variance curve is a graphical representation commonly used in PCA. It illustrates the cumulative amount of variance explained by each principal component as more principal components are added to the analysis.

When creating the code, already existing libraries were used, such as KMeans, PCA, silhouette_score and StandardScaler from sklearn. After importing the dataset from google drive, we created a function that calculates wcss and plots elbow curve. A standard scaler and PCA were implemented in the function to reduce the number of parameters. For the cumulative variance curve, we used normalized data.

It was not possible to accurately determine the number of transmitters from the curves of the elbow method shown in Figure 3. From the charts it would be possible to estimate the number of transmitters at either 4 or 9. As for the differences between the different PCA values, only the scale of the y-axis (WCSS) was different. In terms of shape, the individual courses were almost the same. It follows that the influence of hardware differences was significantly more significant on some parameters then on others.

From the graph (Figure 4), it can be seen that the first 6 components are most significant during PCA. Other components slightly brought the cumulative explained variances closer to 1 and after 9 parameter the course was practically constant.

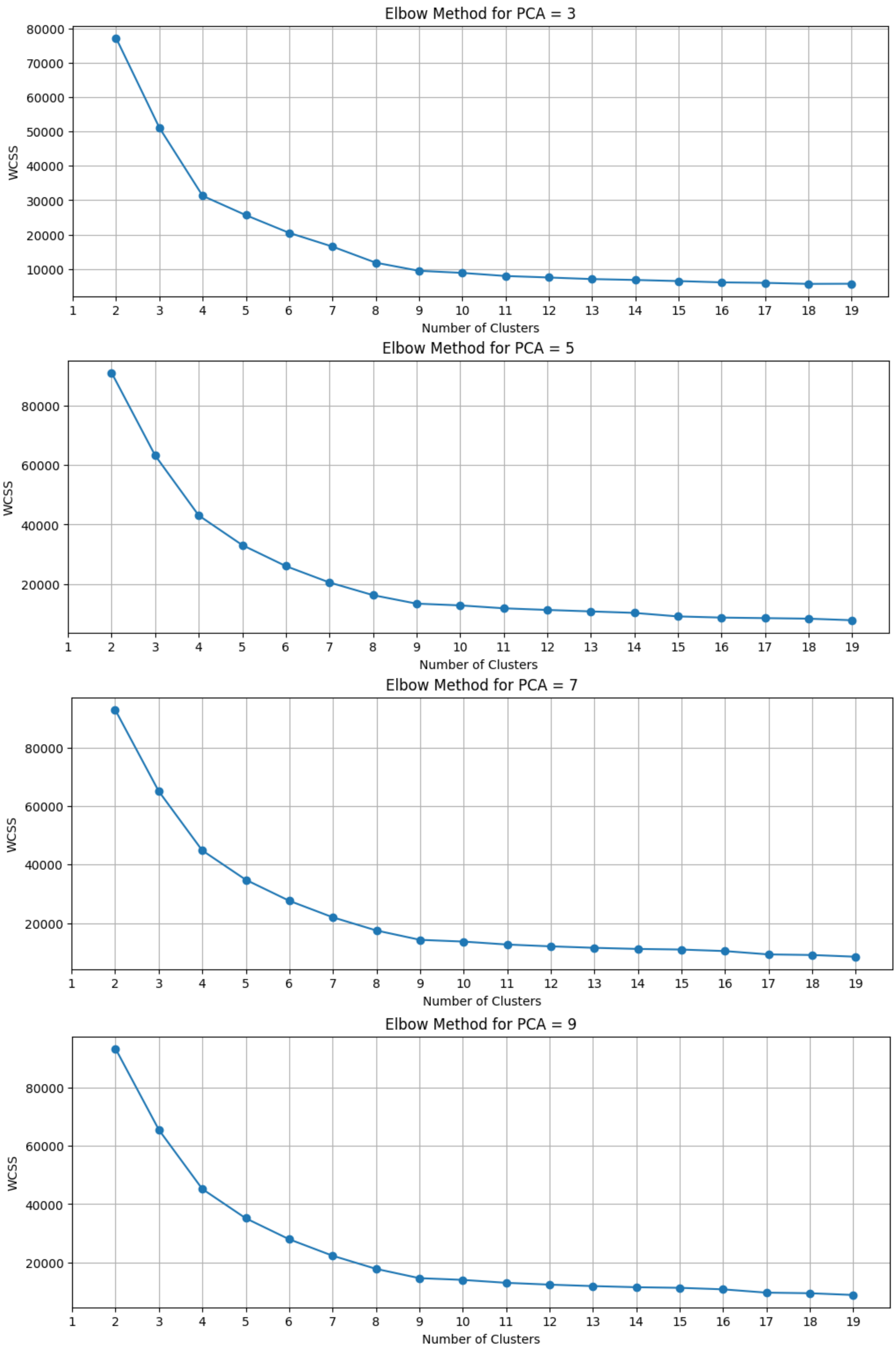

Figure 3: Elbow method for PCA = 3, 5, 7, 9

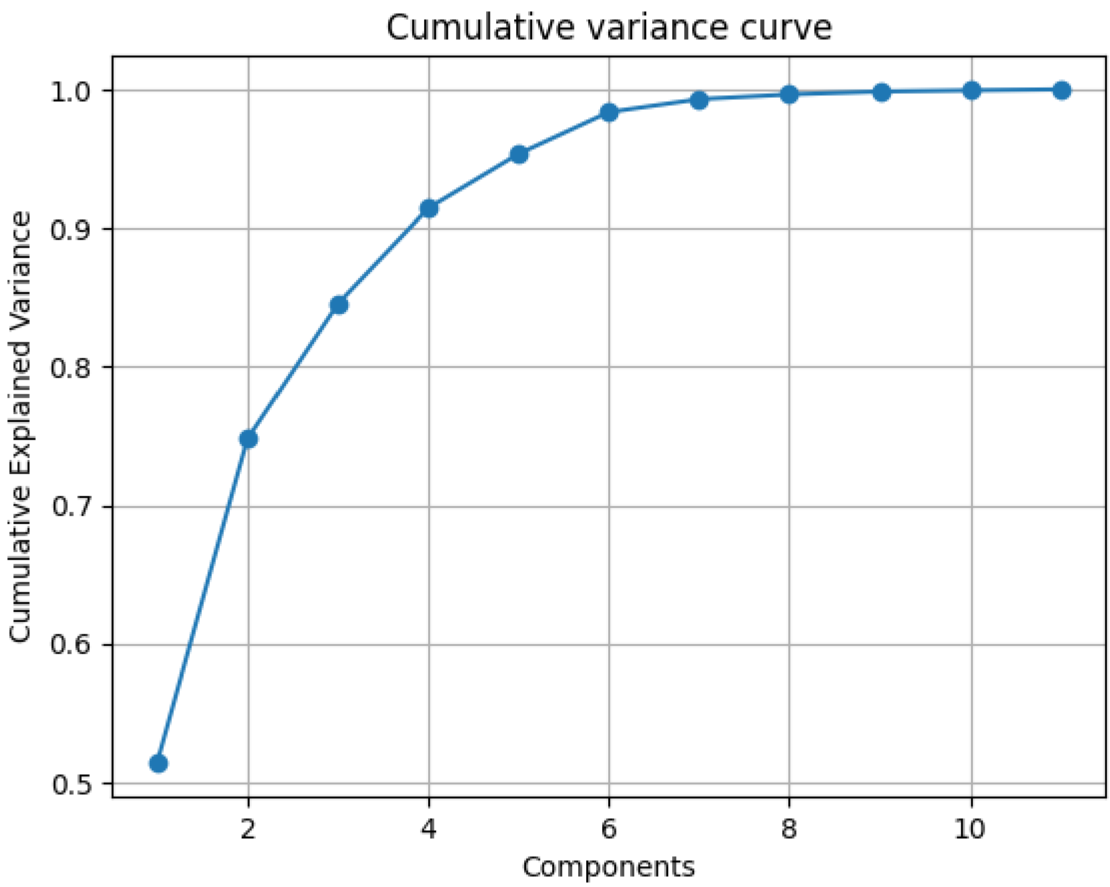

Figure 4: Cumulative variance curve

Since it was not possible to determinate the precise number of transmitters from the elbow method, a silhouette analysis was also implemented.

Silhouette analysis is a method to evaluate the quality of clustering results by measuring how similar each data point is to its own cluster compared to other clusters. A high silhouette score indicates that the data points in a cluster are well-separated from other clusters and that the clustering solution is good.

Silhouette analysis is displayed in Figure 5. Here it is more apparent how many transmitters were used. Silhouette score grew up to the value of groups 9, at 10 was almost the same and after began to decline. From this course, the number of transmitters can be determined at 9 or 10. Considering the previous methods, we assume that the number of transmitters was 9.

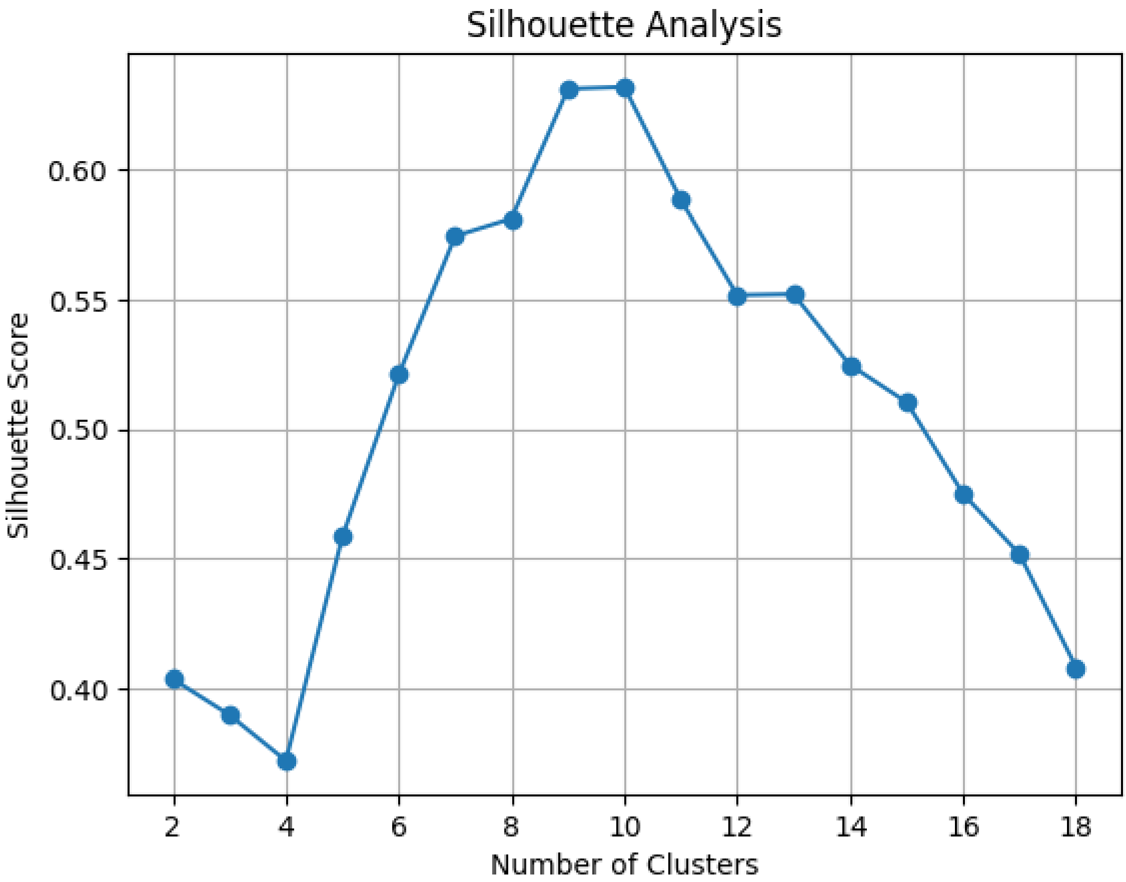

Figure 5: Silhouette analysis

Conclusion

The task of this project is to determinate the number of wireless transmitters based on the hardware imperfections. We utilized two methods to determine the correct value of k (“number of clusters”), namely elbow method and silhouette analysis. Elbow method is shown in Figure 3 for different set up of PCA. After thorough examination, it's theoretically possible to deduce the appropriate number of clusters, but the distinction isn't very clear. Hence, we opted for the Silhouette method. The correct number of clusters can thus be clearly inferred from the Silhouette Analysis (Figure 5). The optimal value is obtained for k = 9."

Github: https://github.com/marekcrn/MLF/blob/main/Project-mini/Project1.ipynb