Question 4 – principal component analysis

For solving this question I have went through the simple steps of calculating the eigenspectru:

* I read data from file
* I also calculated the scatter matrix to check the calculation of eigenvector and eigenvalues but I no longer used it because I used the Matlab libraries for those.
* I calculated the 21x21 covariance matrix and used the *cov*() function from Matlab because I already learnt at assignments how to calculate it and now I wanted to focus more on displaying the results
* I calculated the eigenvalues and eigenvectors using the Matlab function *eig()* which takes as input the covariance matrix that was calculated previously
* I plotted the eigenspectrum by taking the diagonal of matrix eigenvalues which was obtained previously and had size 21x21 but the eigenvalues as placed on diagonal and rest of the matrix is 0. I sorted the eigenvalues on diagonal descending and displayed the results.

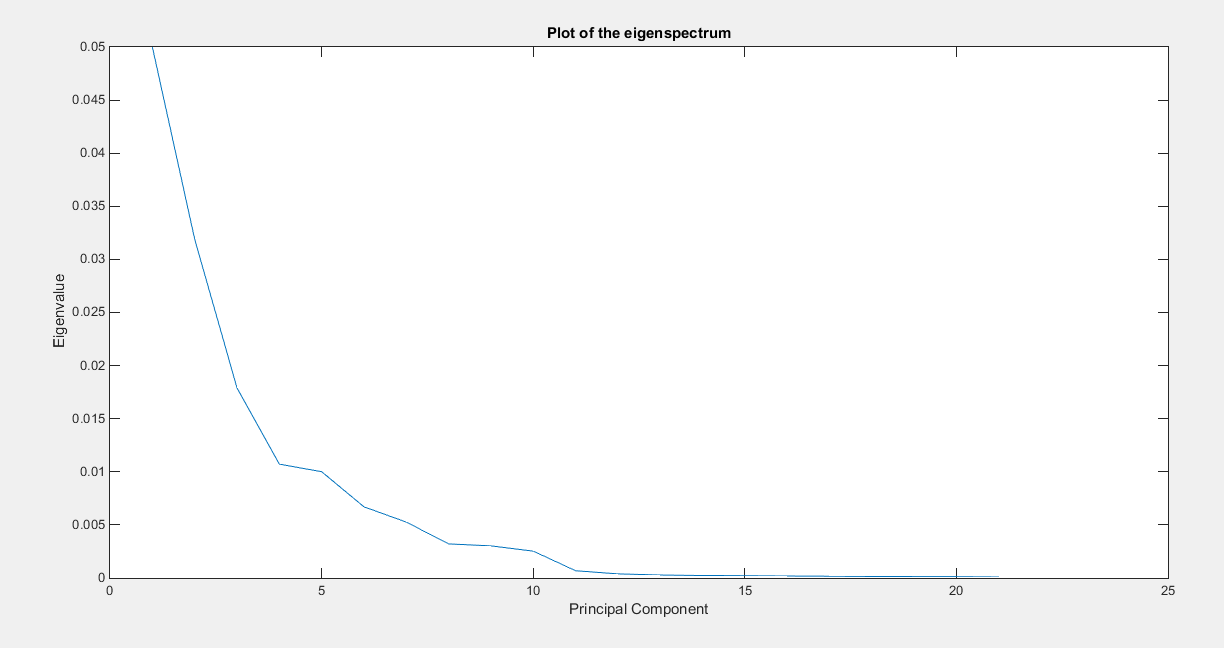


Figure 1 Eigenspectrum for the train data from file keystrokesTrainTwoClass.csv

The plot from above shows that the first 5 parameters have great significance in scatter of the points in space while the other parameters have very low eigenvalue.

For plotting the first two principal components I used Matlab function *eigs* which also calculates the eigenvectors but I can choose to get returned only eigenvectors and eigenvalues for the first two principal components. After calculating the eigenvector for those, I transposed the 21x2 matrix that was obtained and multiplied with the data matrix to transform the 21dimensional points into 2d space. By plotting the first two principal components it should be visible how the patterns and grouped.

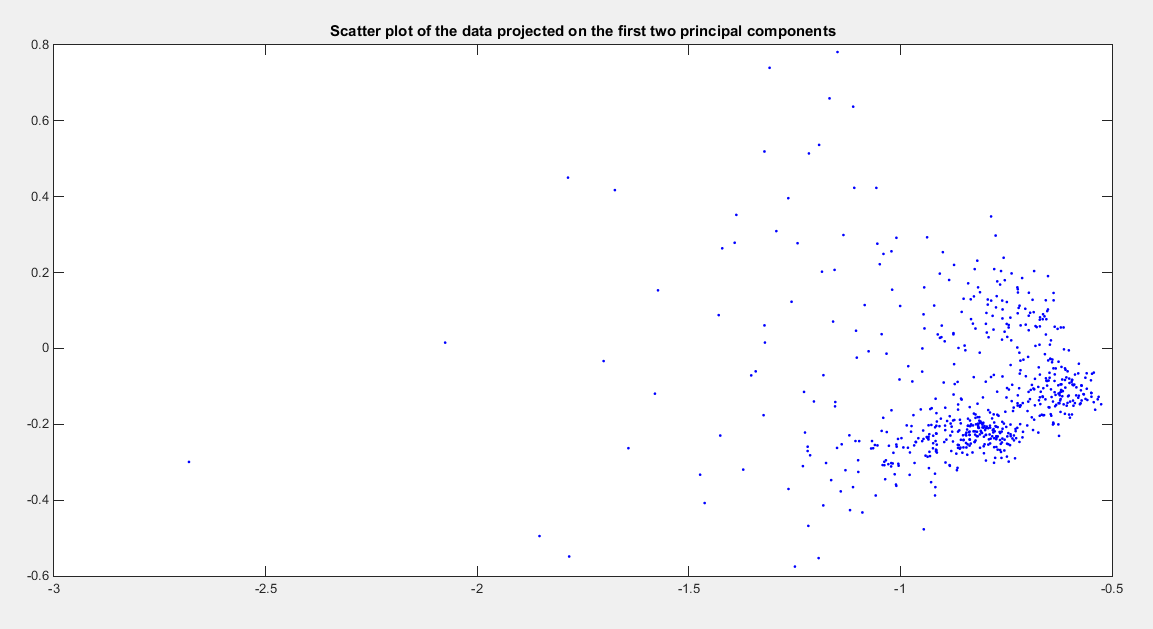


Figure 2. Scatter plot of the data projected on the first two principal components

The scatter plot shows that there are two concentrations of patterns which the most likely correspond to the two types of patterns from data that was used for classification at previous question. Now it is easy to understand the high accuracies for the binary classification algorithms that I have used: the patterns from each type are very well grouped together, so it is easy, for example, to find a linear separator for the train points.