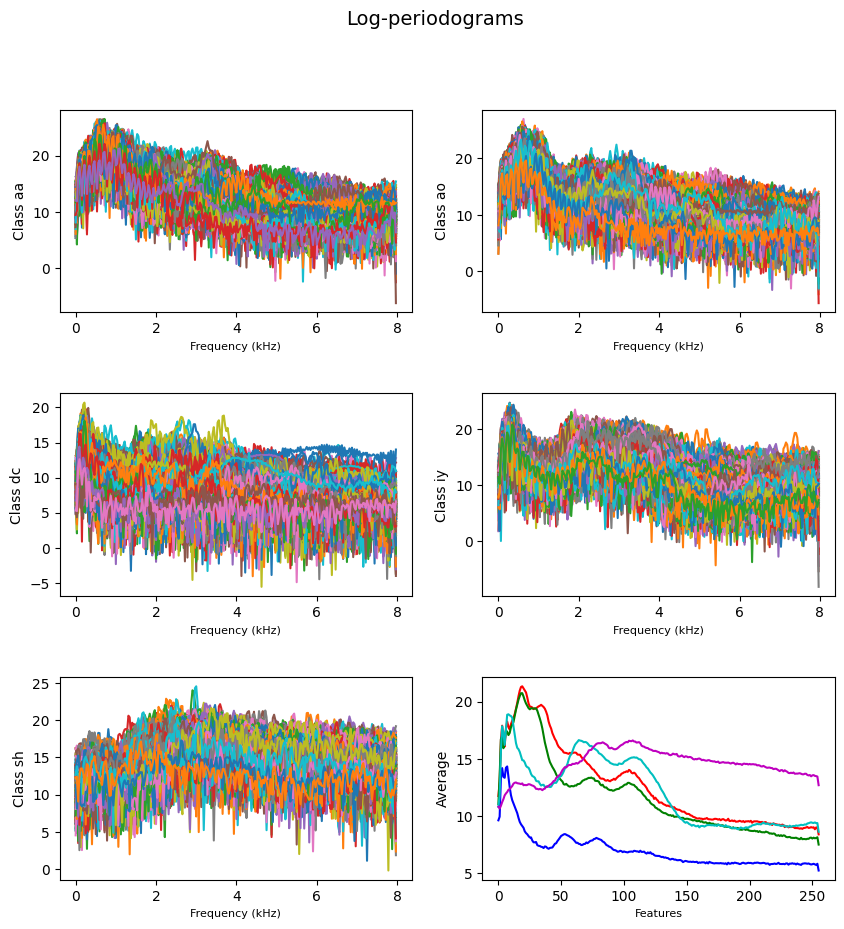
**MACHINE LEARNING FROM DATA**

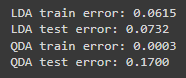
**Report: Lab Session 2 – Feature selection - PCA**

**Names:Abdullah Qureshi, S M Rakib Hasan and Niina Hietamäki**

**Questions**

Q1. Include the plots of the phoneme spectra.

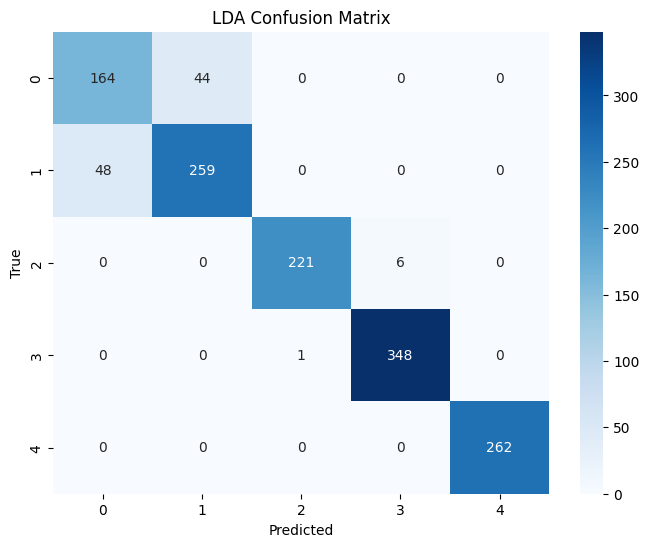
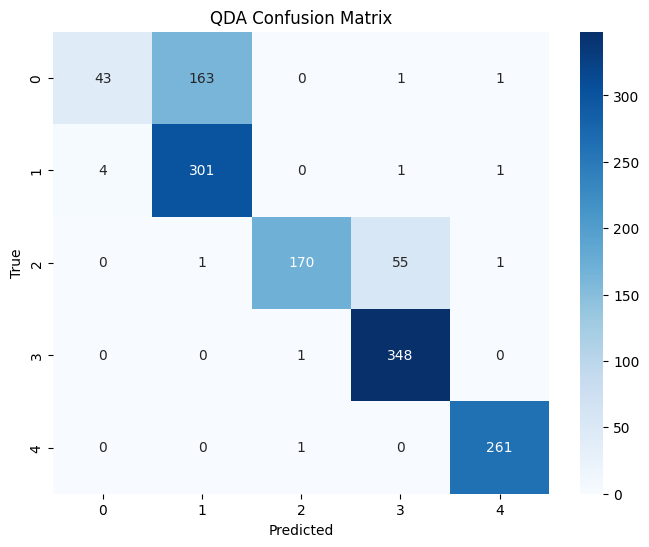
Q2. Include the error probabilities for the training and test sets obtained with the linear classifier (LC) and the quadratic classifier (QC), using all the features. Discuss the results.



LC: Train error : 0.0615 (6.15%) and test error: 0.0732 (7.32%)

QC: Train error : 0.0003 (0.3%) and test error: 0.1700(17%)

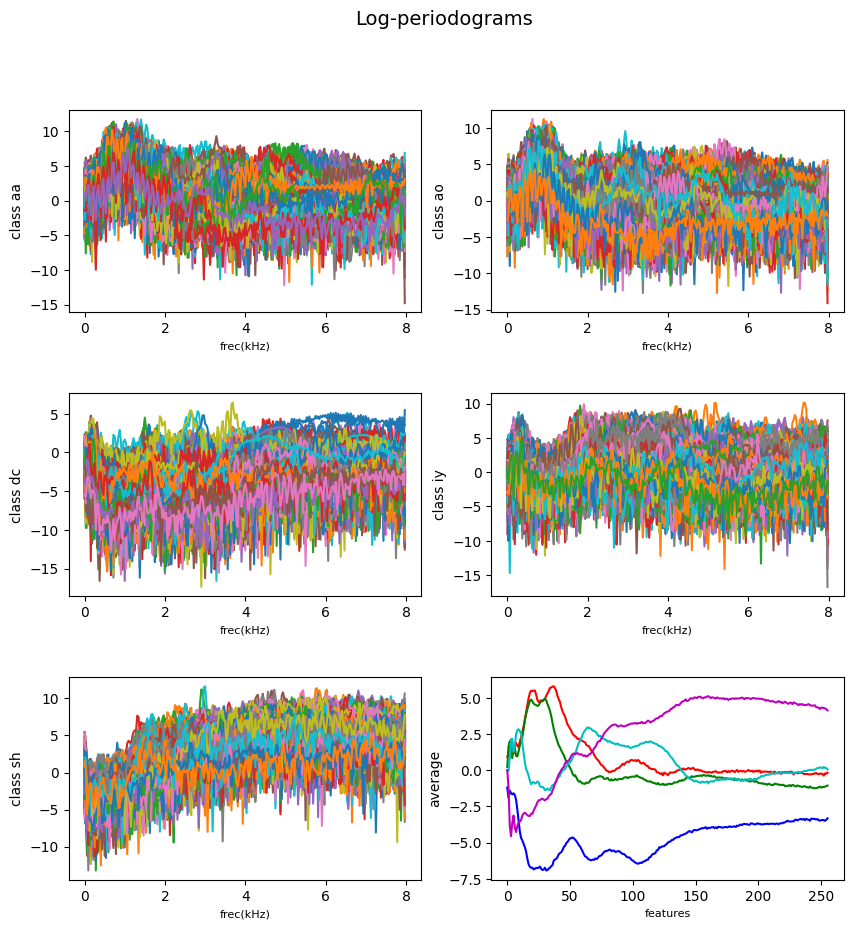
For LC the performance is more balanced. The training and test errors are relatively close to each other, which indicates that the model generalises better to new, unseen data. QC the training error (0.0003) is so small that the model is practically perfectly fitted to the training data. However, this can lead to the model also learning the random noise in the data, which results in poor generalisation to the test data , as reflected by high test error (0.17001)

Q3. Include the confusion matrices for the test set obtained with the linear classifier (LC) and the quadratic classifier (QC), using all the features. Discuss the results.

LC: Performs generally better than QC, especially classes 0 and 1. The classifications are more balanced and generalises well across most classes.

QC: Has significantly large issues, especially with classes o and 2, which indicates that it is sensitive to the mixing of certain classes. It is likely overfitted to the training data, which is reflected in poor performance on the test data.

Q4. Which features would you choose? Show the error probabilities for the training and test sets obtained with the linear and the quadratic classifier. Compare with the previous case (using all features) and discuss the results.



Using the spectrum plots to try to find the most discriminative features we observe that around 105 and 75 values the classes look distinguishable from each other. If you use only these features for training the linear and the quadratic classifier, we get these results

LDA test error: 0.478197

QDA test error: 0.467110

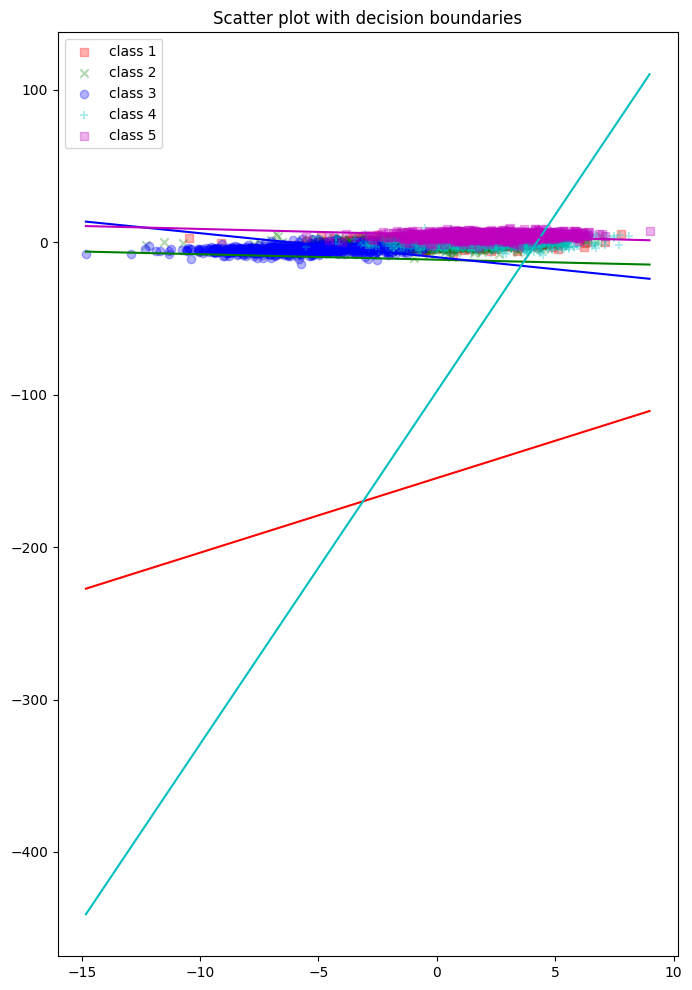
Now choosing another set of random features [60, 135] we get the following results

LDA test error: 0.392461

QDA test error: 0.382853

It is clear that reducing the dimensionality and number of features reduce the accuracy of both the Linear and Quadratic classifiers in this case. Using all the features we got a LC error of 0.0732 (7.32%) and QC error of 0.1700(17%).

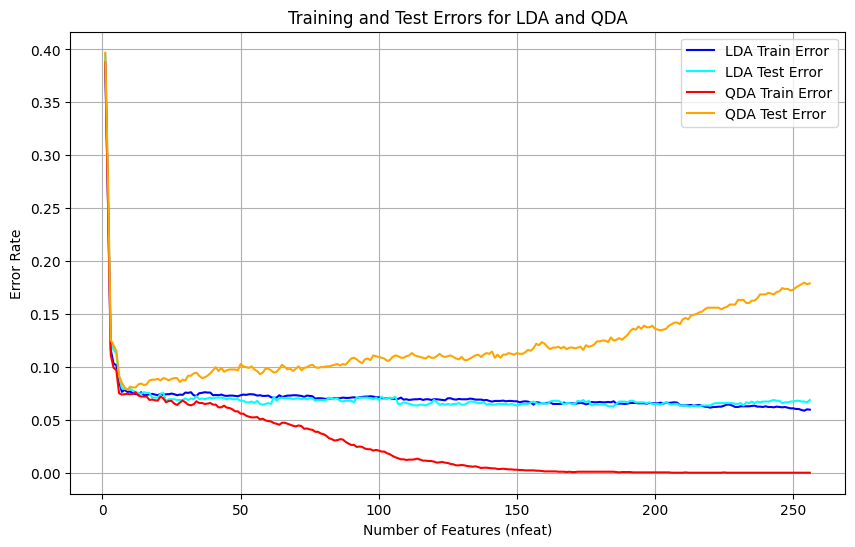
Q5. Include the scatter plot and decision boundaries obtained. Discuss the results.



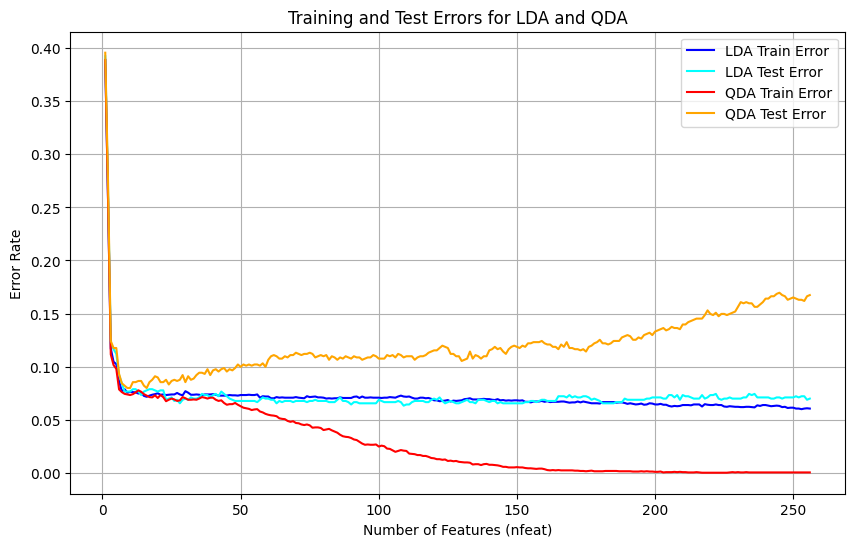
This scatter plot shows that the number of features selected are not enough to draw clear decision boundaries between different classes. For example the class 4 will be misclassified most of the time due to very unclear decision boundary. The variance explained by each of the selected components is not enough resulting in large errors. We need higher dimensionality and add more features to get an accurate LC or QC classifier.

Q6. Show the error curves for the linear and the quadratic classifier on the training and on the test set.

Answer: The error curves for the linear and the quadratic classifier on the training and on the test set with 70/30 split.



The error curves for the linear and the quadratic classifier on the training and on the test set with 80/20 split.



Q7. Discuss which dimension is the most adequate for the linear classifier and which is the best one for the quadratic classifier. Remember that it is important not to overfit on the training data (the test error should not be much larger than the training error).

Answer: For a 80/20 split, the optimum dimension for lda is at 103 dimensions.

for the qda, it is at 15 dimensions. Code:

