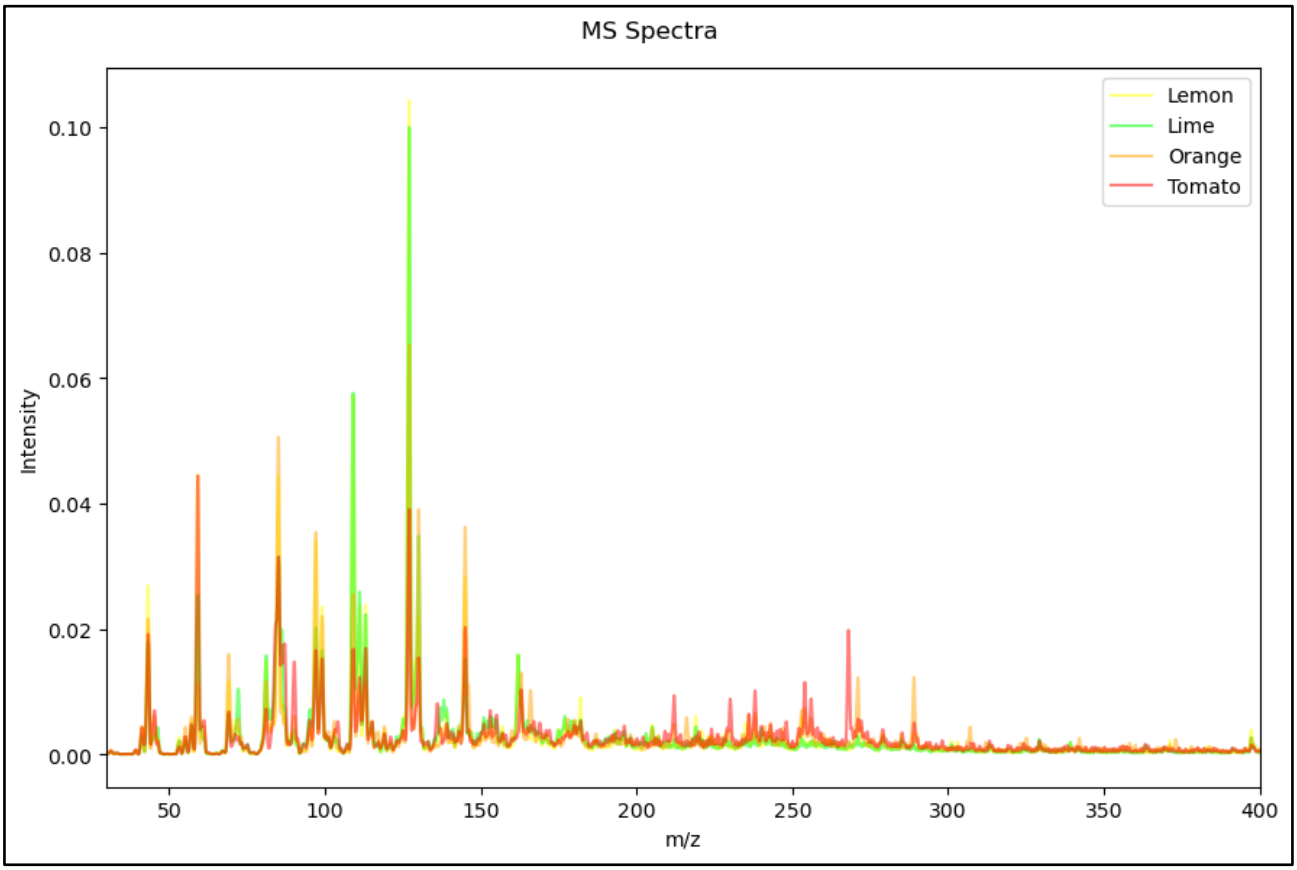
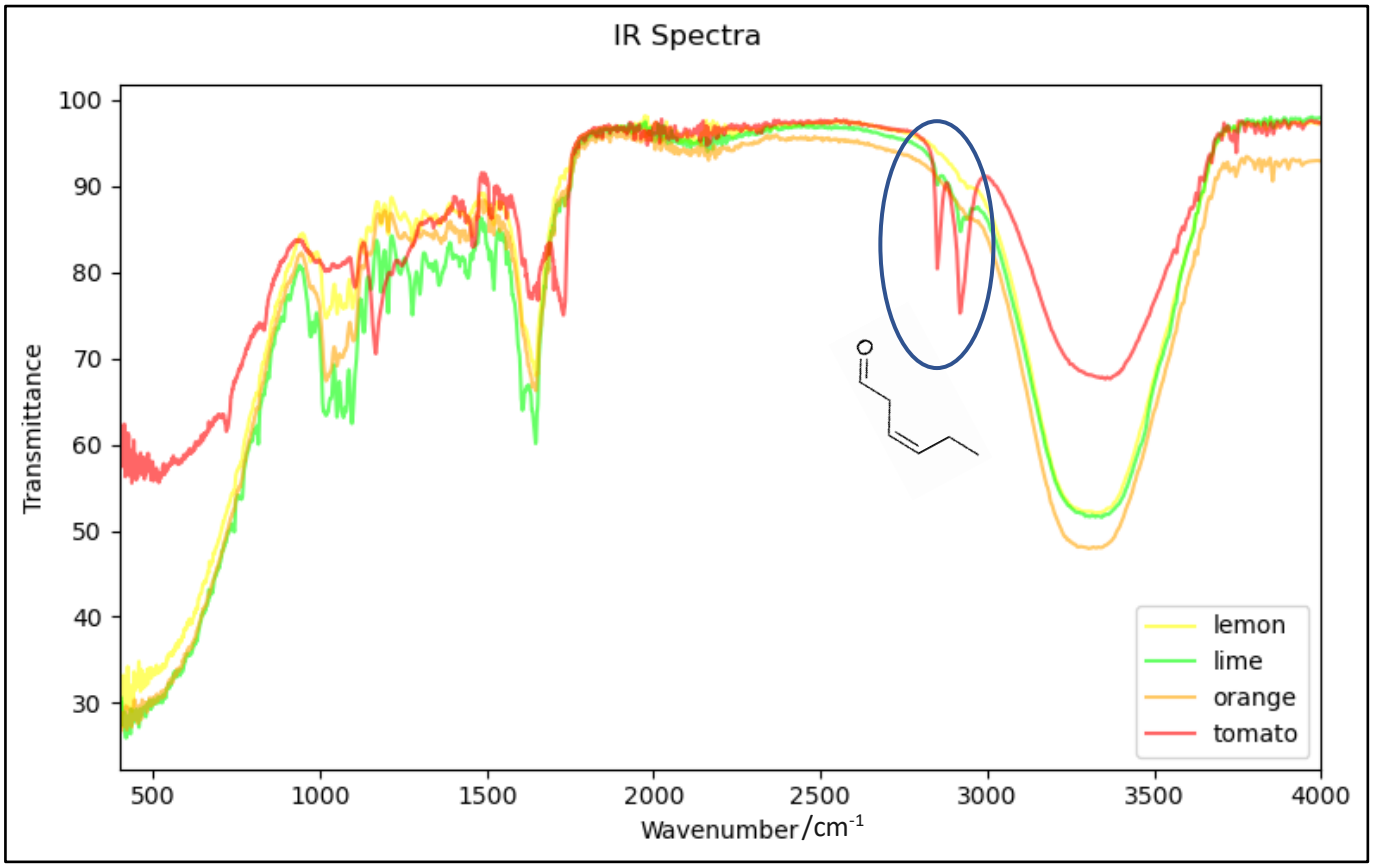
CHEM6164 Project – Classifying MS and IR Spectra of Fruits

Background

Analytical sciences form an integral part of drug-development, medical diagnostics, and fields beyond life sciences such as forensics and agricultural chemistry. These fields heavily rely on understanding quantitatively and qualitatively the composition of mixtures, which they achieve using techniques such as nuclear magnetic resonance, chromatography, electron microscopy, and a range of spectroscopic techniques [1] including mass (MS), X-ray, and vibrational spectroscopy (IR). With the advent of high-throughput chemistry [2] and better data management [3], the size of chemical databases has grown substantially over the past decades. As a result, the output of analytical techniques, i.e., spectra, can no longer be analysed by humans at a reasonable pace and scale. Machine learning-based (ML) methods [4] are promising tools for accelerating each step of the analysis of chemicals [1], including data acquisition, noise reduction, data reconstruction, labelling and classification of data. By automating the workflow of, for instance, medical imaging [5], environmental, research and development costs can be substantially lowered whilst preventing the need for intrusive surgeries, illness, and avoidable deaths. Unsupervised clustering techniques have been shown to accelerate the profiling of volatile organic compounds from deconvolved GC/MS spectra [6], which is highly relevant to drug-development. Inorganic material design has also benefitted from techniques such as graph-based neural networks which can be trained to predict X-ray absorption near-edge spectra [7]. ML and dimensionality reduction techniques can also be used to identify complex and overlapping patterns which are concealed from the human eye, as did Pérez et al. [8] to the 1H NMR spectra of metabolomics samples. Xu et al. [9] utilised deep learning to differentiate between typically developing children from children with autism spectrum disorder whilst gaining a better understanding of the underlying causes of autism. Whilst some of these models possess extraordinary predictive capabilities, scalability and transferability often limits their application domain. Techniques such as transfer learning [10] can overcome this because they require less data and computational power for training. Explainability is also a key challenge in ML, especially for applications that use ‘black-box’ models [11] such as neural networks, since fields like medical diagnostics require transparent and bias-free models.





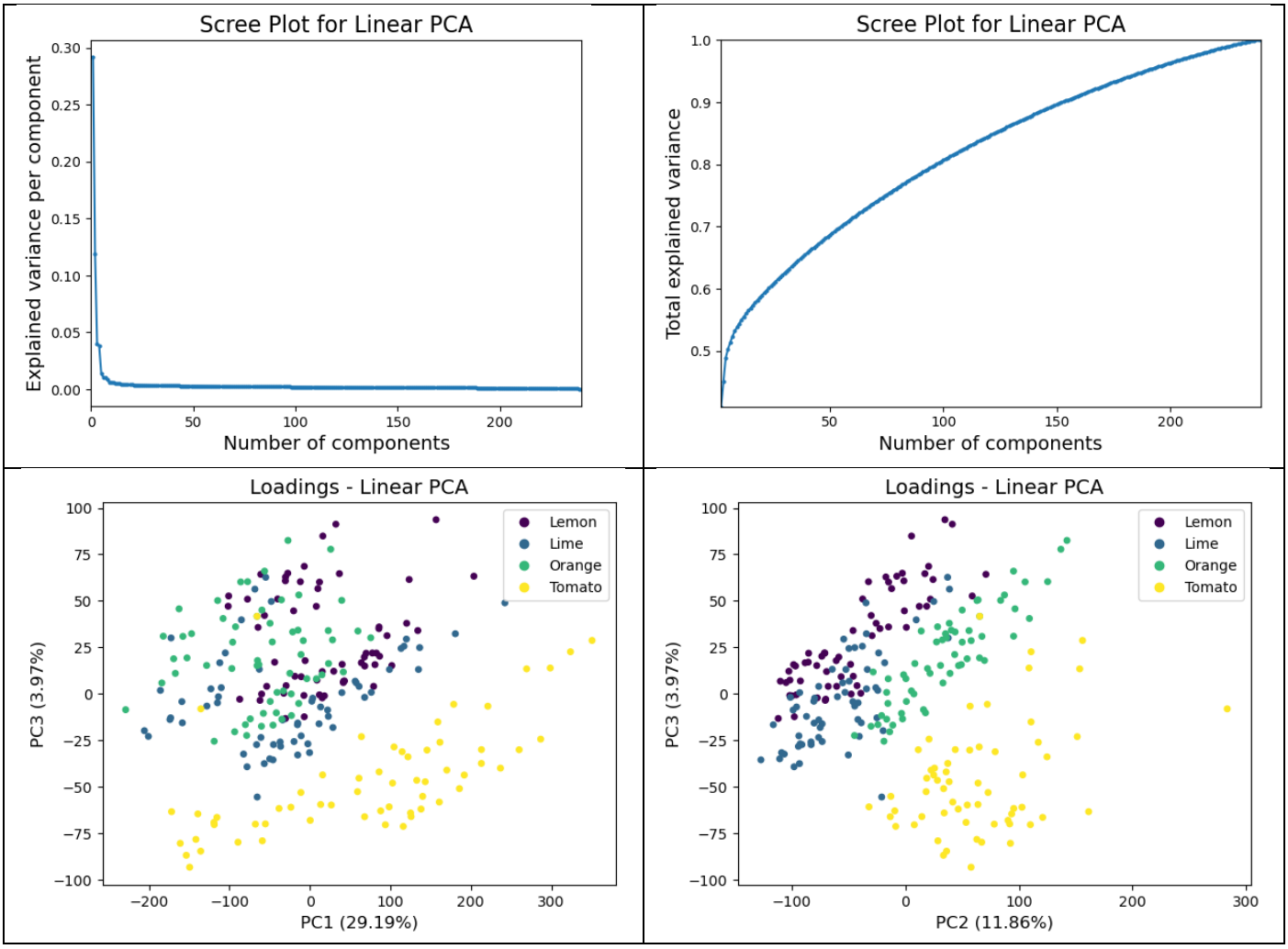
Exploratory Data Analysis

Across both IR and MS spectra (see above) citrus fruits (lemon, lime, and orange) exhibited many overlapping peaks, making their classification by visual inspection challenging. In the MS spectra, tomato showed significantly more intense peaks above 200 m/z which can be traced back to carotenoids – high mass, conjugated polyenes. Citrus fruits, on the other hand, yielded peaks mostly accumulating below 150 m/z, with various intensities. Limes exhibited notably more intense peaks around 120 m/z. The FTIR spectra showed again three, largely analogous spectra for the citrus fruits. Since lemon and orange each contains a distinct enantiomer of limonene, their spectra were almost identical. Lime only differed from them by a noticeable peak splitting at 1600 cm^{-1} and a minor but sharp peak around 2900 cm^{-1}. Tomato's IR spectrum produced less peaks, with two large stretching frequencies at 2700 and 2900 cm^{-1} (circled), most likely due to the vibrations of the aldehyde group of cis-3-Hexenal, a volatile substance responsible for the aroma of ripe tomatoes [12]. Citrus fruit exhibited more intense vibrations around 1600 cm^{-1}, signalling that citrus fruits have more aromatic content than tomatoes.

Dimensionality Reduction

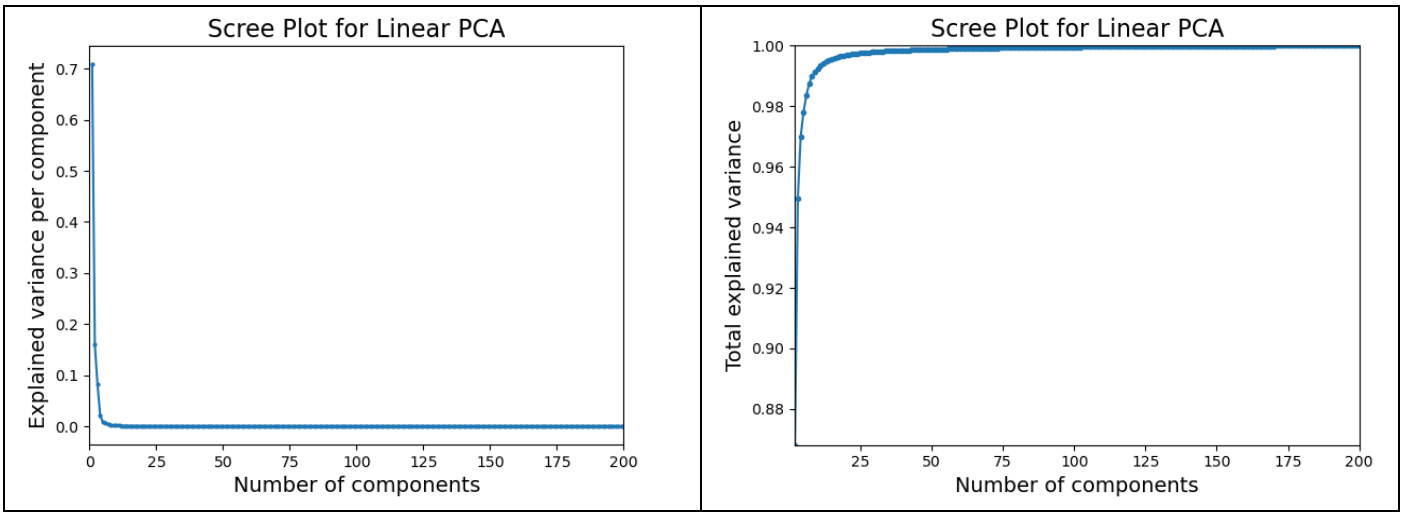
Mass Spectrometry Data

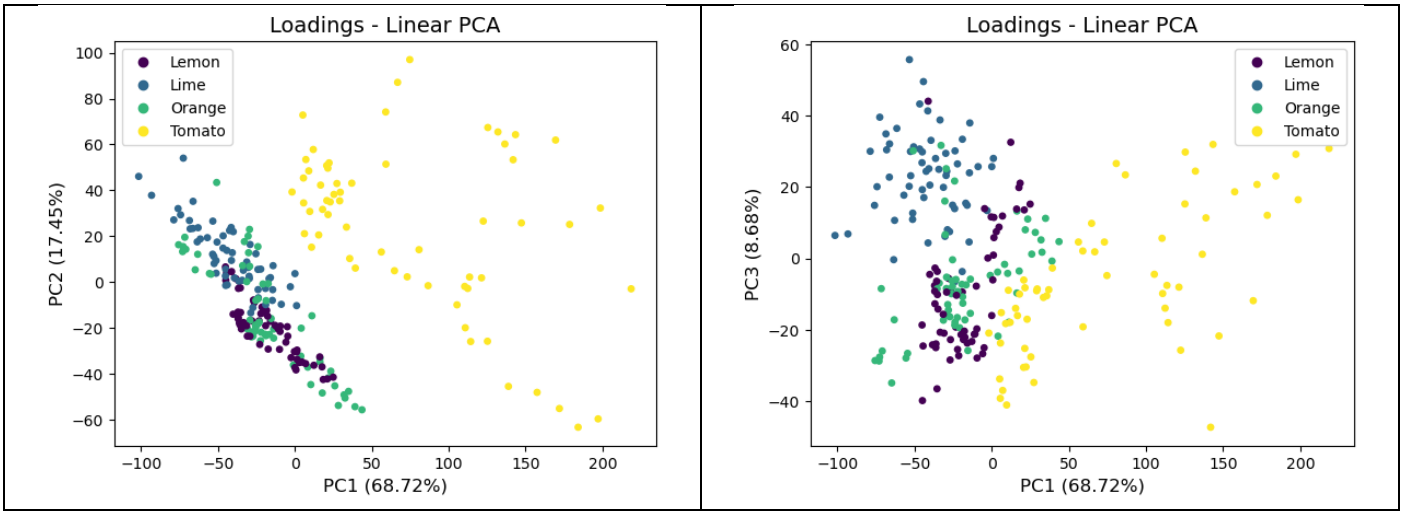
After merging all data and scaling them to have a mean of zero and variance of one, linear PCA was performed to reduce the dimensionality of the datasets. The first five principal components (PCs) explained 50% of the MS dataset, after which adding PCs increased variance by small margins. For sufficient predictive power, 150 PCs were selected which described 90% of the variance – exact values of variance as as function of the number of PCs can be found in the Appendix. The first three PCs were visualised using 2D scatterplots. No distinct clusters formed along PC1 (see Appendix) whilst the second and third PCs separated tomatoes out very well. The citrus fruits, especially lemon and lime formed overlapping clusters with PC3 distinguishing orange somewhat from them.



Infrared Spectroscopy Data

Almost all variance was explained by the first 10 components of the transformed IR data, for which reason 10 PCs were selected for future work. The first PC separated tomato from the citrus fruits, and the combination of the first and third PCs distinguished lemon and lime well. Orange's spectrum overlapped with lemons, which was expected due to the stereochemical relationship that links the two fruits.





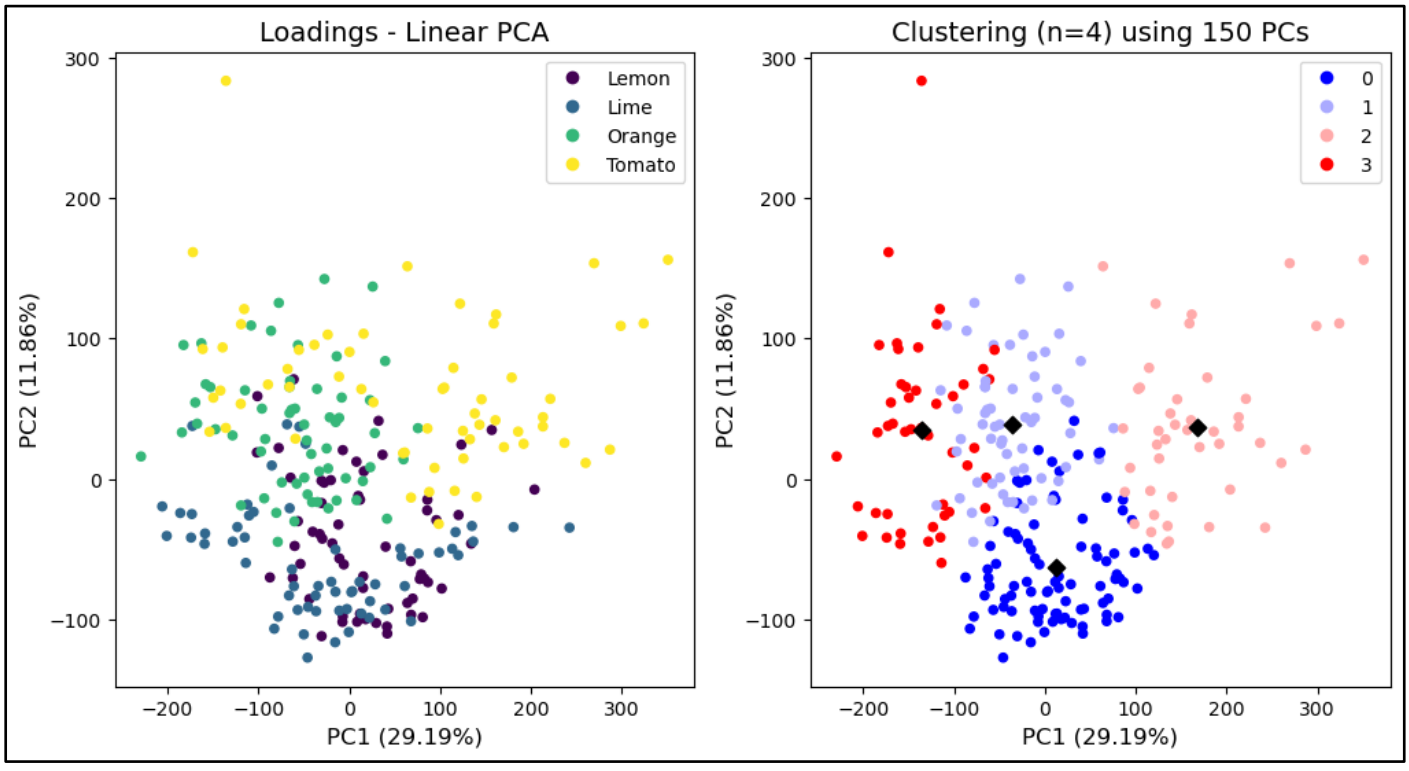
Comparison

The IR data was sparser than the MS because the first 10 components described almost the entire dataset. In both types of spectra, tomato formed well-separated cluster, despite the substantially different variance captured by the first three PCs. Citrus fruit formed overlapping clusters, with some segmentations.

Unsupervised Learning

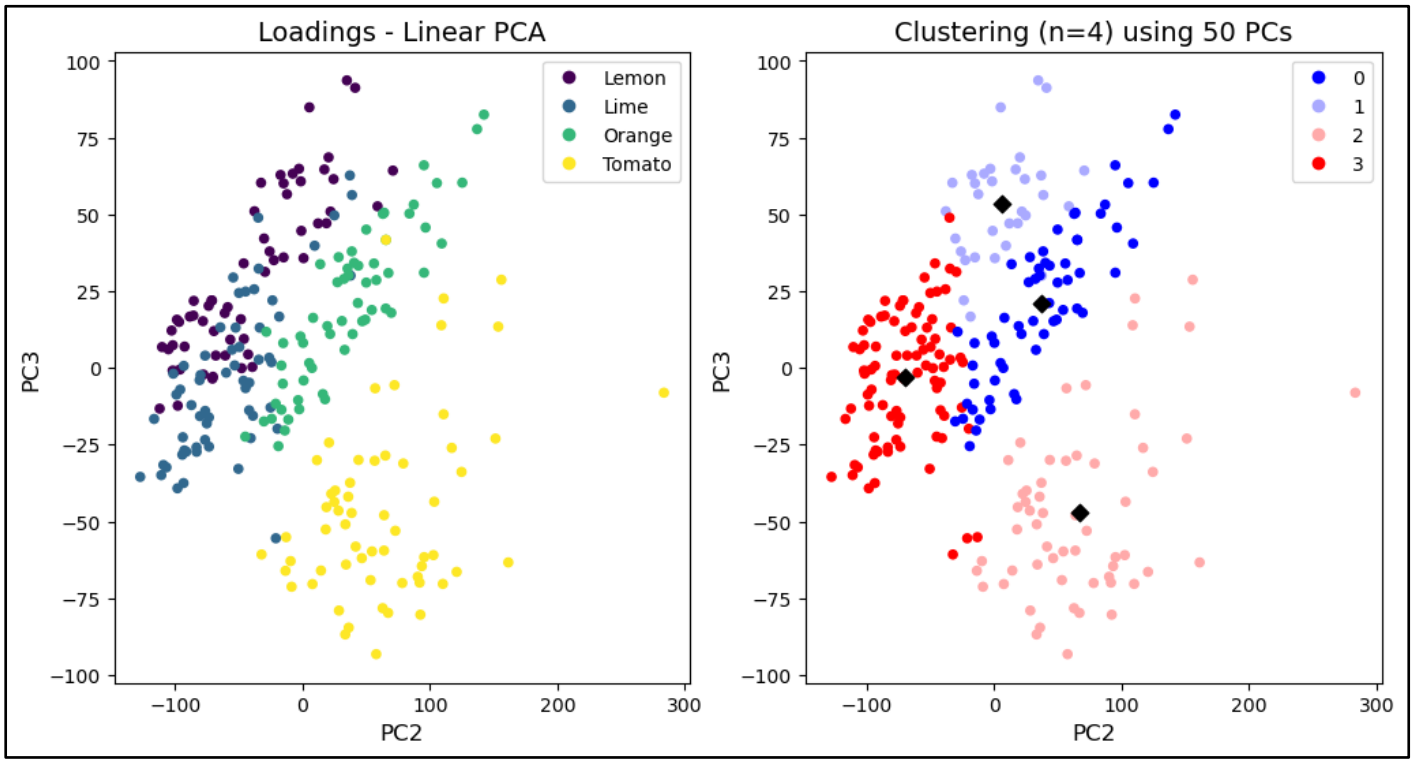
Mass Spectrometry Data

K-means clustering was done initially with four clusters (n=4) using 150 PCs. The clusters formed failed to classify the data well as the algorithm struggled to differentiate between limes and lemons, and to find the boundaries between fruits in general.



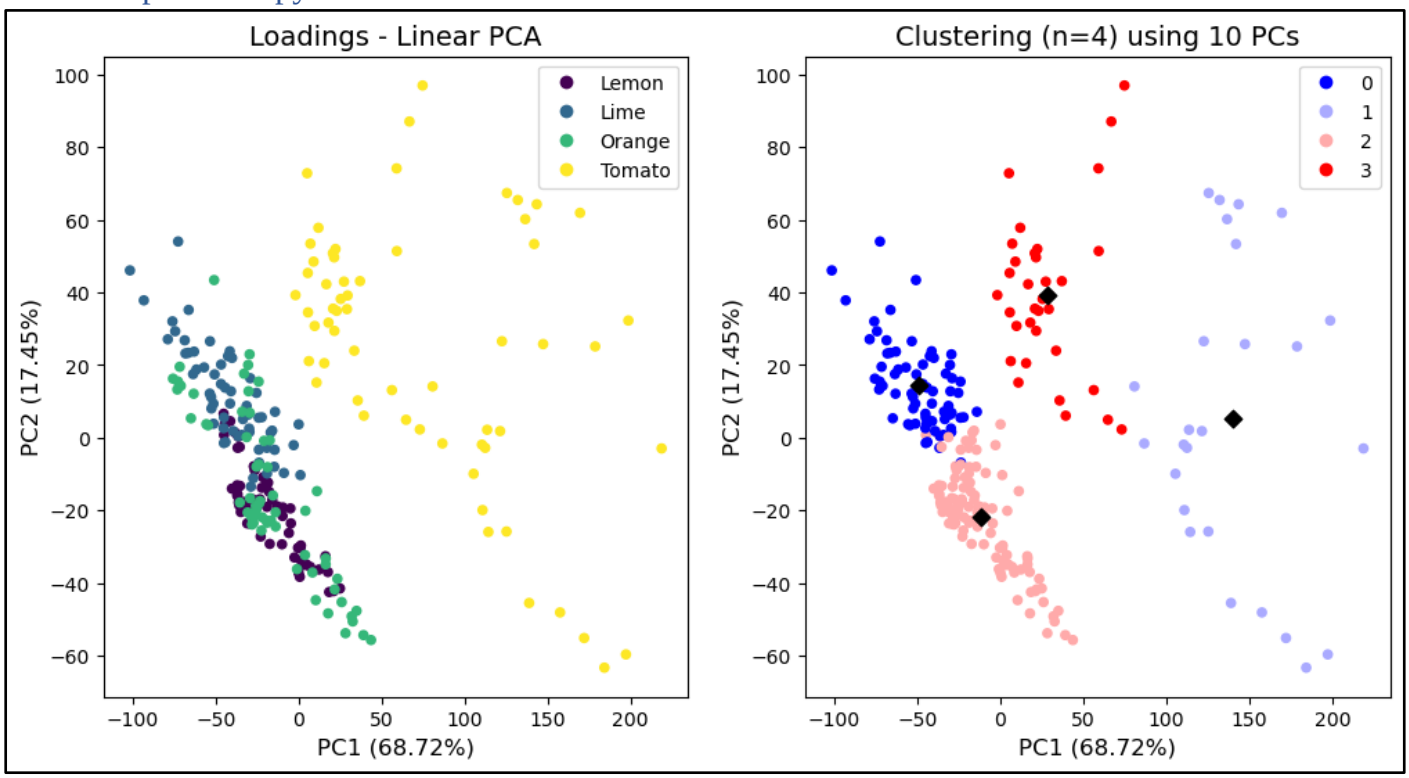
Reducing the number of PCs improved the Silhouette score for the clustering, however when visualised, no observable difference was found in the scatterplots as can be seen in the Appendix. This was probably due to

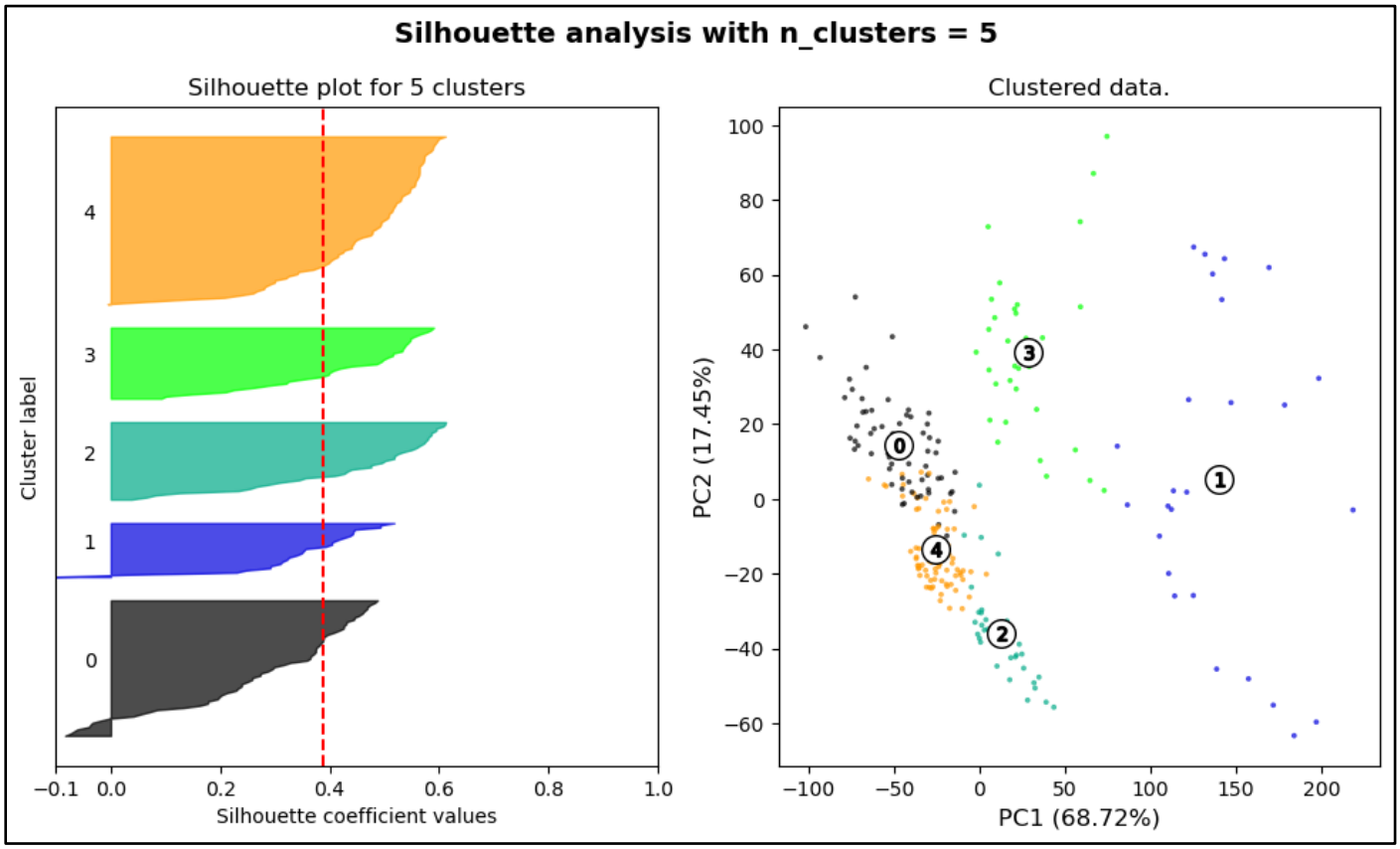
the majority of the PCs explaining only a small fraction of the MS data. Interestingly, re-doing the clustering in the absence of the first PC (see below) classified almost all oranges and tomatoes correctly, even when less PCs were used.



As the next step, the number of predefined clusters were changed to investigate natural cluster formations, however none of the newly formed clusters or their combinations aligned with the actual fruit classes. Silhouette scores can be found in the Appendix.

Infrared Spectroscopy Data





Performing k-means clustering with n=4 did not distinguish classes well for the IR data neither, especially not for citrus fruits. Increasing the number of principal components lowered the average Silhouette score, however not as drastically as for the MS data (see Appendix). Upon varying the number of predefined clusters, the average Silhouette score remained near 0.4, however visualising n=5 (see above) showed that the clusters (1) and (3) together accounted for all the tomato samples, and the remaining three classes are distributed between the three citrus fruits, improving the prospects of classification using clustering.

Supervised Learning

Success Metrics

To assess the performance of the support vector classifiers, the following metrics were selected: accuracy, F1 score, recall, and precision. Some of the models were evaluated with confusion matrices as well.

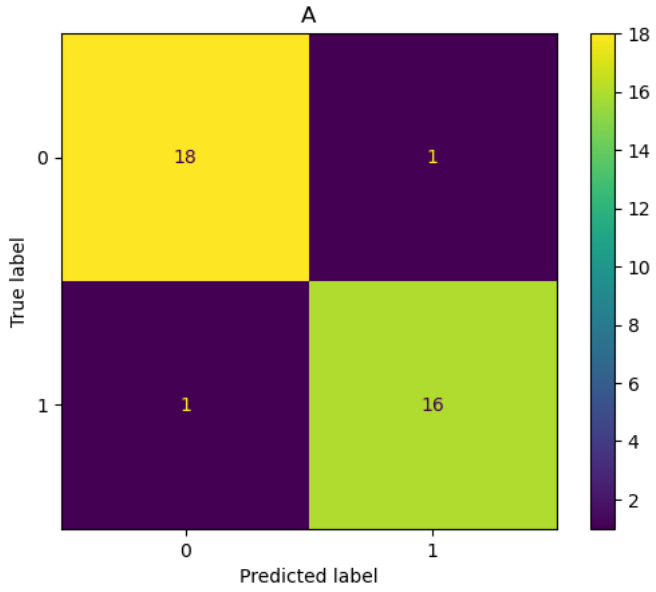
Mass Spectrometry Data

For the separable and non-separable pair, the tomato-orange and lemon-lime pairs were selected.

The support vector machine (SVM) classified the tomatoes and oranges with ease. Regardless of the value of C or the size of the train set, the SVM yielded a perfect accuracy and F1 score of 1.0 on the test set. The RBF kernel, however, classified all test samples into one class, despite the large range of gamma values tested.

The SVM was able to classify the non-separable lemon-lime pair with high performance. Lower C values produced a slightly higher accuracy of 0.94 (see Appendix), only misclassifying one sample per class. Unsurprisingly, a smaller training set resulted in slightly lower performance however due to the random selection of training samples, results may vary per re-run. Decreasing the number of PCs in most cases resulted in lower performance as was expected, especially for model trained using less than five PCs. As the number of PCs decreased, the accuracy decreased mostly due to a lower recall, meaning an increase in the number of false positives (higher number of lemons misclassified as limes).

Confusion matrix for the SVM model using C=0.01



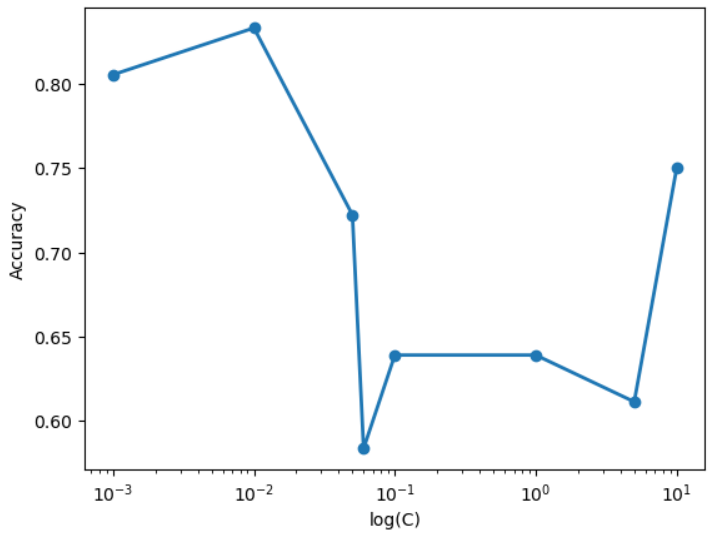
Performance of the SVM with a varying number of training samples (top) and PCs (bottom)

#Training samples	Accuracy	F1 score	Precision	Recall
36.0	0.916667	0.923077	0.9	0.947368
60.0	0.944444	0.944444	1.0	0.894737
84.0	0.916667	0.914286	1.0	0.842105

	#PCs	Accuracy	F1 score	Precision	Recall
0	2.0	0.472222	0.000000	0.000000	0.000000
1	3.0	0.694444	0.645161	0.833333	0.526316
2	5.0	0.916667	0.923077	0.900000	0.947368
3	10.0	0.972222	0.972973	1.000000	0.947368
4	20.0	1.000000	1.000000	1.000000	1.000000
5	50.0	0.916667	0.918919	0.944444	0.894737
6	100.0	0.916667	0.923077	0.900000	0.947368

Infrared Spectroscopy Data

The tomato-orange and lemon-orange pairs were selected as the separable and non-separable, respectively. The separable case again produced immaculate accuracy and F1 scores of 1.0 regardless of the value of C, the number of train samples or the number of PCs used. The RBF kernel yielded the same results for gamma values as low as 0.0001, however above that it classified all test samples into one class as for the MS data.



Effect of C on accuracy (left), and performance with varying training samples (right, top) and varying #PCs (right, bottom) for the non-separable case.

#Training samples	Accuracy	F1 score	Precision	Recall
36.0	0.944444	0.947368	0.900000	1.0
60.0	0.972222	0.972973	0.947368	1.0
84.0	0.972222	0.972973	0.947368	1.0

	#PCs	Accuracy	F1 score	Precision	Recall
0	1.0	0.416667	0.400000	0.411765	0.388889
1	2.0	0.611111	0.681818	0.576923	0.833333
2	3.0	0.583333	0.705882	0.545455	1.000000
3	4.0	0.527778	0.679245	0.514286	1.000000
4	5.0	0.555556	0.692308	0.529412	1.000000
5	6.0	0.555556	0.692308	0.529412	1.000000
6	7.0	0.555556	0.692308	0.529412	1.000000
7	8.0	0.555556	0.692308	0.529412	1.000000
8	9.0	0.750000	0.800000	0.666667	1.000000
9	10.0	0.833333	0.857143	0.750000	1.000000

The non-separable case produced varied results. Upon increasing the value of C, accuracy decreased until C=10 where it jumped 6% below the highest accuracy. With increasing C, precision tended to increase whilst recall often decreased. This suggests that a larger C, that is a harder SVM margin yields less false positives (in this case oranges classified wrongly as lemons) but more false negatives (lemons misclassified as orange). Reducing the size of the training set decreased accuracy driven by lower precision. This in combination with a stable recall suggests that less training examples lead to a higher number of false positives – oranges

misclassified as lemons. Switching the linear kernel to RBF worked reasonably well, producing an accuracy of 0.81 with gamma values below 0.001 (see Appendix). More interesting was the dependence of the linear model on the number of PCs. Given that almost the entire IR dataset was explained by the first 10 PCs, accuracy drastically fell with less for less than 10 PCs, mostly due to lower precision, i.e., more oranges misclassified as lemons.

Comparison

For both MS and IR data, the separable classes consistently yielded accuracy and F1 scores of 1.0, however the RBF kernel struggled to tackle linearly separable data with ease. For the non-separable classes, accuracy largely depended on the number of PCs used, and in some cases the RBF kernel proved better than the linear SVM. In comparison with the k-means clustering, the SVM was able to tackle both linearly separable and non-separable classification and offered more options for tuning hyperparameters.

Conclusion

PCA was used to transform 60 MS and 60 IR spectra of four fruits, reducing their dimensions by up to two magnitudes. k-means clustering and classification using SVMs was performed on the transformed data, using easily separable and non-separable pairs of fruit as use cases. The k-means clustering was in general mislead by the first PC of each dataset and could not label overlapping classes correctly. For better cluster segmentation, PCA with non-linear kernels such as polynomial or RBF kernel could be explored. The SVMs produced very high accuracy and F1 scores for both easily separable and non-separable fruit-pairs but were less immune to changes in the number of PCs used. The best combination of hyperparameters (kernel, C, #PCs, gamma) could be found using gradient search but also alternative algorithms such as tree-based models or neural networks – especially convolutional neural networks – are worth exploring. The neural network-based models however require substantially more data and therefore, for a task like this, the use of transfer learning [10] may be the only viable option. Local explanation methods [13] such as SHAP [14] could reveal the decision-making mechanism of the SVM classifiers and aid the work of experimental chemists in the future who face a similar task of classifying fruits.

References

(1) R. Houhou and T. Bocklitz, *Anal. Sci. Adv.*, 2021, **2**, 128–141.
(2) N. S. Eyke, B. A. Koscher and K. F. Jensen, *Trends Chem.*, 2021, **3**, 120–132.
(3) M. D. Wilkinson et al. *Sci. Data*, 2016, **3**, 1–9.
(4) Aurélien Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition*, O'Reilly Media, Inc., 2019.
(5) C. A. Peña-Solórzano, D. W. Albrecht, R. B. Bassed, M. D. Burke and M. R. Dimmock, *Forensic Sci. Int.*, DOI:10.1016/j.forsciint.2020.110538.
(6) Y. Alkhalifah et al. *Anal. Chem.*, 2020, **92**, 2937–2945.
(7) R. Houhou and T. Bocklitz, *Anal. Sci. Adv.*, 2021, **2**, 128–141.
(8) Y. Pérez, M. Casado, D. Raldúa, E. Prats, B. Piña, R. Tauler, I. Alfonso and F. Puig-Castellví, *Anal. Bioanal. Chem.*, 2020, **412**, 5695–5706.
(9) L. Xu, Y. Liu, J. Yu, X. Li, X. Yu, H. Cheng and J. Li, *J. Neurosci. Methods*, 2020, **331**, 108538.
(10) K. Weiss, T. M. Khoshgoftaar and D. Wang, *J. Big Data*, 2016, **3**, 9.
(11) R. Guidotti, A. et al. *ACM Comput. Surv.*, DOI:10.1145/3236009.
(12) R. G. Buttery, R. Teranishi and L. C. Ling, *J. Agric. Food Chem.*, 1987, **35**, 540–544.
(13) C. Molnar, *Interpretable Machine Learning*, 2nd edn., 2022.
(14) S. M. Lundberg, G. G. Erion and S.-I. Lee, *arXiv Prepr. arXiv1802.03888*.

Appendix

Variance explained by PCA

MS data

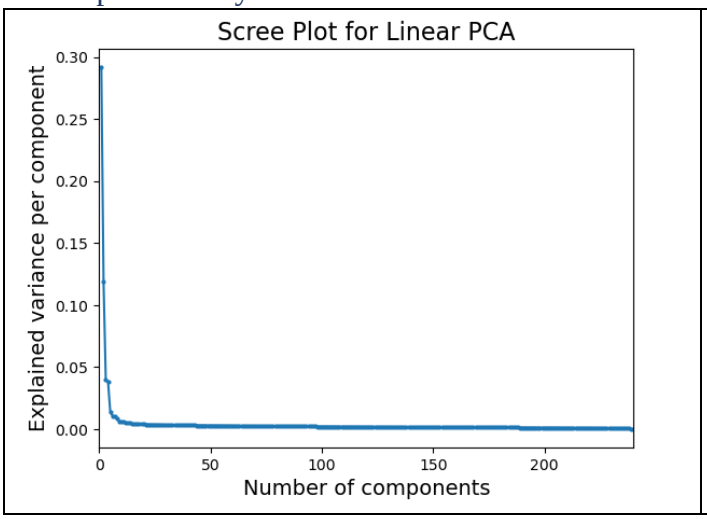
	n_components	total_var
0	2.0	0.410434
1	3.0	0.450109
2	5.0	0.502678
3	20.0	0.589355
4	50.0	0.682823
5	100.0	0.802405
6	150.0	0.892982
7	200.0	0.963411
8	240.0	1.000000

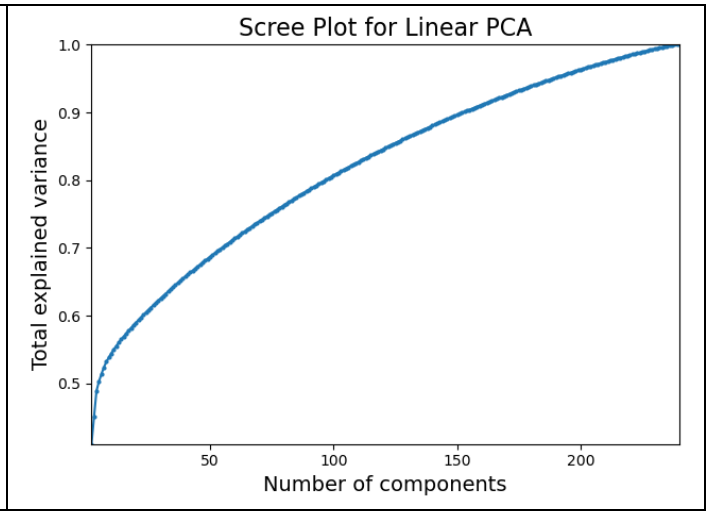
IR data

	n_components	total_var
0	2.0	0.861678
1	3.0	0.948471
2	5.0	0.976681
3	10.0	0.991968
4	20.0	0.996393
5	30.0	0.997600
6	50.0	0.998497
7	100.0	0.999393
8	150.0	0.999762
9	200.0	0.999935

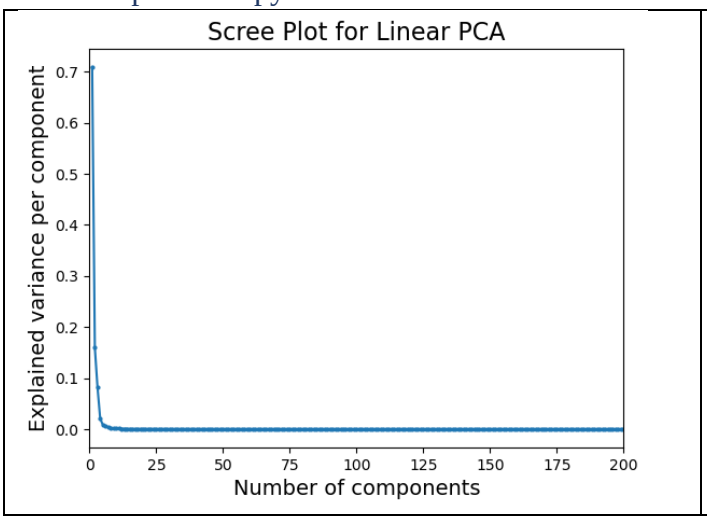
Scree plots

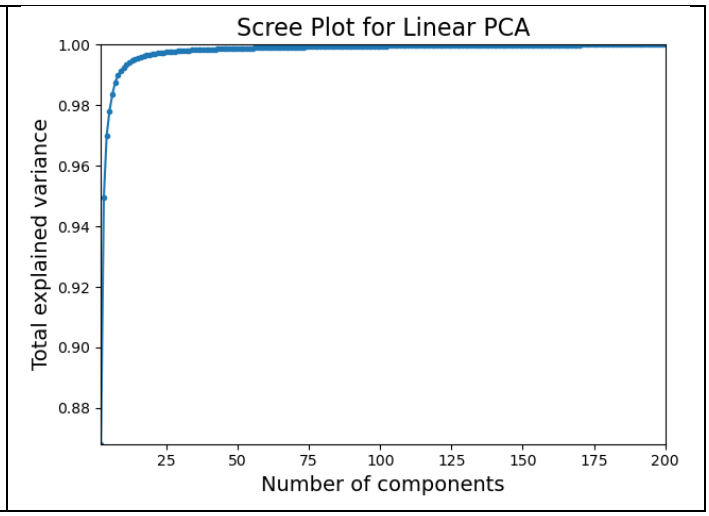
Mass Spectrometry Data





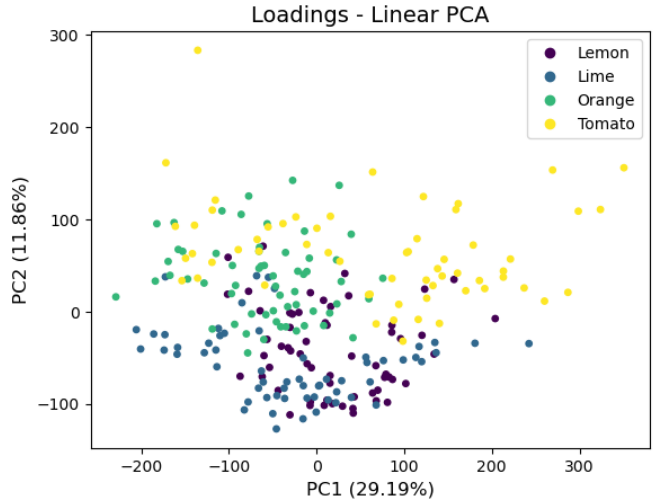
Infrared Spectroscopy Data



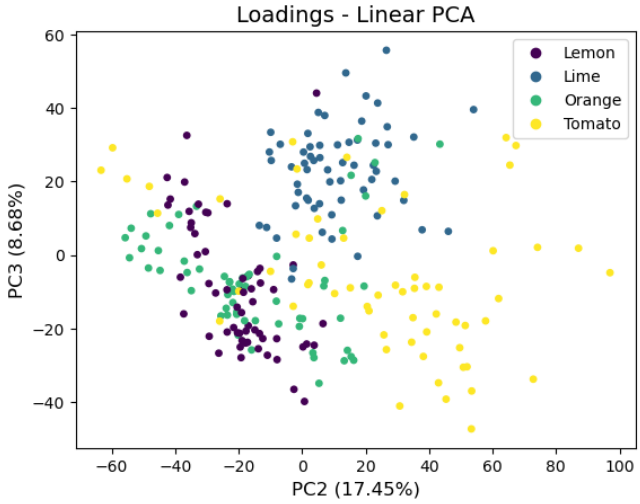


Additional loading plots

Mass Spectometry data



Infrared Spectroscopy data



Clustering and PCs

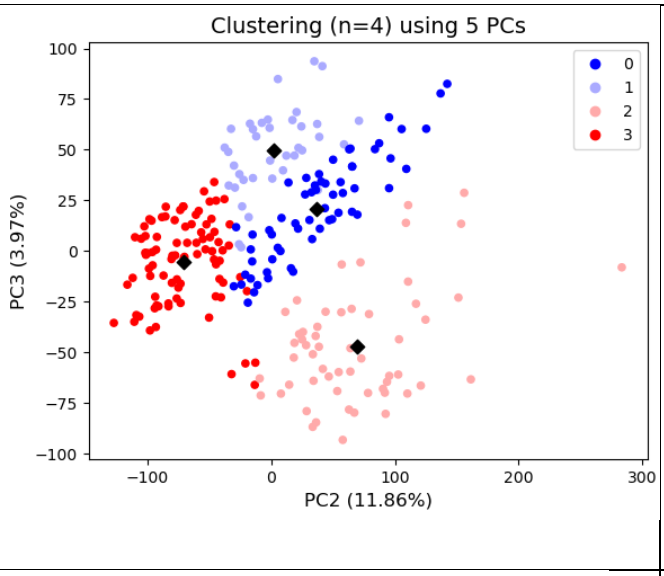
Mass Spectrometry Data

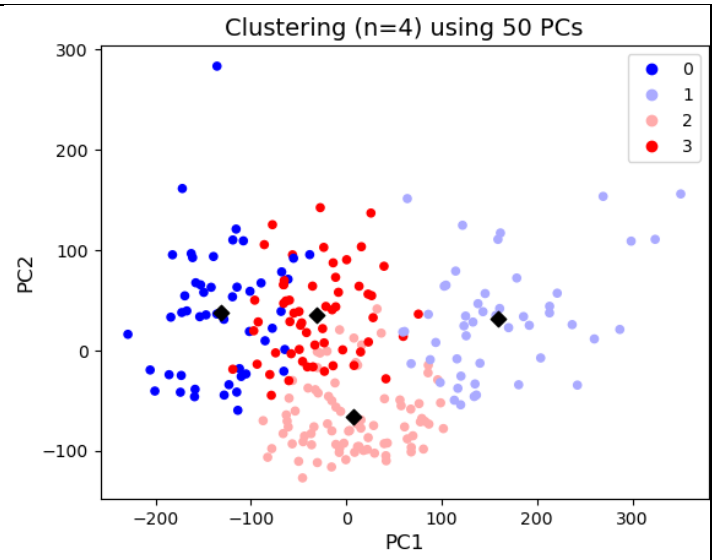
All PCs

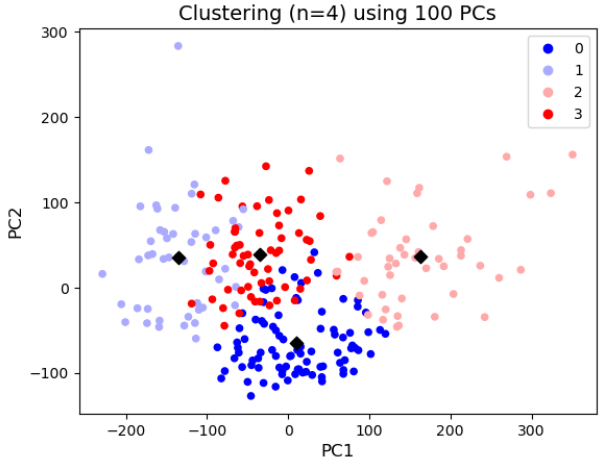
	PCs	Mean Silhouette coeff
0	2.0	0.378039
1	3.0	0.326142
2	5.0	0.309503
3	10.0	0.274999
4	50.0	0.202806
5	100.0	0.155246
6	150.0	0.132635
7	200.0	0.111059
8	240.0	0.106886

First PC taken out

	PCs	Mean Silhouette coeff
0	1.0	0.584897
1	2.0	0.453195
2	4.0	0.399256
3	9.0	0.344776
4	49.0	0.218833
5	99.0	0.157587
6	149.0	0.120462
7	199.0	0.104659
8	239.0	0.095380







Clustering was not affected by a change in the number of PCs

Infrared Spectroscopy Data

	PCs	Mean Silhouette coeff
0	2.0	0.466410
1	3.0	0.416625
2	5.0	0.399102
3	10.0	0.384812
4	20.0	0.379945
5	30.0	0.378454
6	50.0	0.377327
7	100.0	0.376167
8	150.0	0.375186
9	200.0	0.375431

	k	Mean Silhouette coeff
0	2.0	0.583359
1	3.0	0.316049
2	4.0	0.384812
3	5.0	0.392006
4	6.0	0.390929

Performance metrics for the clustering.

Supervised learning

Mass Spectrometry Data

- Separable: tomato-orange pair

	C	Accuracy	F1 score	Precision	Recall
0	0.001	1.0	1.0	1.0	1.0
1	0.100	1.0	1.0	1.0	1.0
2	0.500	1.0	1.0	1.0	1.0
3	1.000	1.0	1.0	1.0	1.0
4	3.000	1.0	1.0	1.0	1.0

#Training samples	Accuracy	F1 score	Precision	Recall
36.0	1.0	1.0	1.0	1.0
60.0	1.0	1.0	1.0	1.0
84.0	1.0	1.0	1.0	1.0

Gamma	Accuracy	F1 score	Precision	Recall
0.000001	0.472222	0.641509	0.472222	1.0
0.000010	0.472222	0.641509	0.472222	1.0
0.000100	0.472222	0.641509	0.472222	1.0
0.001000	0.472222	0.641509	0.472222	1.0
0.010000	0.472222	0.641509	0.472222	1.0
0.100000	0.472222	0.641509	0.472222	1.0
1.000000	0.472222	0.641509	0.472222	1.0
5.000000	0.472222	0.641509	0.472222	1.0
10.000000	0.472222	0.641509	0.472222	1.0

- Non-separable: lime-lemon pair

	C	Accuracy	F1 score	Precision	Recall
0	0.001	0.944444	0.947368	0.947368	0.947368
1	0.010	0.944444	0.947368	0.947368	0.947368
2	0.050	0.944444	0.947368	0.947368	0.947368
3	0.060	0.916667	0.918919	0.944444	0.894737
4	0.100	0.916667	0.918919	0.944444	0.894737
5	1.000	0.916667	0.918919	0.944444	0.894737
6	5.000	0.916667	0.918919	0.944444	0.894737
7	10.000	0.916667	0.918919	0.944444	0.894737

Infrared Spectroscopy Data

- Separable: pair
- Non-separable: pair

	C	Accuracy	F1 score	Precision	Recall
0	0.001	0.805556	0.837209	0.720000	1.000000
1	0.010	0.833333	0.850000	0.772727	0.944444
2	0.050	0.722222	0.642857	0.900000	0.500000
3	0.060	0.583333	0.347826	0.800000	0.222222
4	0.100	0.638889	0.480000	0.857143	0.333333
5	1.000	0.638889	0.434783	1.000000	0.277778
6	5.000	0.611111	0.416667	0.833333	0.277778
7	10.000	0.750000	0.689655	0.909091	0.555556

	Gamma	Accuracy	F1 score	Precision	Recall
0	0.0001	0.805556	0.837209	0.720000	1.000000
1	0.0010	0.805556	0.820513	0.761905	0.888889
2	0.0100	0.500000	0.000000	0.000000	0.000000
3	0.1000	0.500000	0.000000	0.000000	0.000000
4	1.0000	0.500000	0.666667	0.500000	1.000000
5	3.0000	0.500000	0.666667	0.500000	1.000000
6	5.0000	0.500000	0.666667	0.500000	1.000000
7	10.0000	0.500000	0.666667	0.500000	1.000000