# Classification of NIRS Data

***Dr Carole Wright***
***Department of Agriculture and Fisheries, Mareeba, QLD, Australia***

## Introduction:

There are numerous supervised classification techniques which have the purpose of assigning objects to different groups. Three of the more commonly used techniques for near-infrared spectroscopy (NIRS) data are:

- **principal components discriminant analysis** (PC-DA)
- **support vector machines** (SVM)
- **partial least squares discriminant analysis** (PLS-DA)

This study compared these three classification methods applied to three NIRS data sets.

## Classification Methods:

**PC-DA** produces a number of orthogonal discriminant functions that minimise the variance within groups, while maximising the separation between groups. Often the number of wavelengths contributing to each spectra is greater than the number of samples. To overcome the requirement that there are more samples than variables, the data dimensionality is reduced using PC analysis prior to applying the DA.

**SVM** is a technique which has the ability to model linear and non-linear classification problems by identifying a hyperplane that maximises group separation.

**PLS-DA** uses PLS regression to discriminate between multiple groups by using an exclusive binary coding scheme to identify group association.

The correct classification rate for each group and the overall correct classification rate was used as the criterion to compare the three methods.

## Data Sets:

The data sets analysed by the three classification methods were:

- Live **mud crabs** were graded based on meat fullness.

*Meat fullness grading system*

| Grade | % Meat Yield |
|---|---|
| **A** | > 45 |
| **B** | 35 – 45 |
| **C** | < 35 |

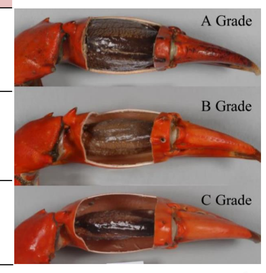
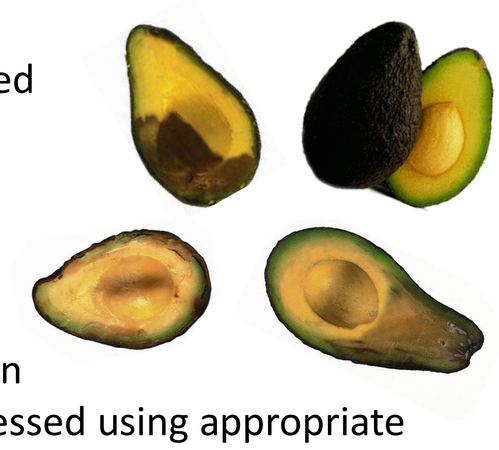



Image: Mayze et al (2016) Final Report FRDC Project 2014/218

- **Avocado** fruit at eating ripe were classed as bruised or not bruised.
- **Avocado** fruit were classed based on rot %: <10% or ≥10%.



In each data set, one spectra represents an individual sample. Spectra were pre-processed using appropriate methods. Independent validation sets were used to assess the models for the two avocado data sets, but due to smaller sample sizes in the mud crab data set, full cross validation was used.

## Results:

- PC-DA (7 PC) was the better overall performer for the 3 grade mud crab data set.
- PC-DA (6 PC) and PLS-DA (5 latent variables) produced the same rate of correct classifications for the avocado bruising which was higher than SVM.
- All three classification methods produced similar overall correct classification rates for the avocado rot data set.

*Avocado validation set and mud crab cross validation number of correct classifications. Highest correct classifications are in bold. n = sample size.*

| Data Set | Group | n | PC-DA | SVM | PLS-DA |
|---|---|---|---|---|---|
| Mud Crab | Grade A | 13 | **10** | 8 | 4 |
| | Grade B | 29 | **23** | 18 | 16 |
| | Grade C | 52 | 44 **(81.9%)** | **49** (79.8%) | 45 (69.1%) |
| Avocado | Bruised | 59 | **49** | 46 | **49** |
| | Not bruised | 57 | **55** **(89.7%)** | 49 (81.9%) | **55** **(89.7%)** |
| | Rot <10% | 113 | **85** | 84 | 82 |
| | Rot ≥10% | 118 | 108 (83.5%) | 111 **(84.4%)** | **112** (84.0%) |

## Conclusion:

These results have shown there is **no one size fits all** classification method, therefore more than one method should be considered.

## Acknowledgments:

The author would like to thank Dr Brett Wedding and Mr Steve Grauf for collecting the spectra and for permission to use the data.



Queensland Government