# Kendall Brown r0773111 Chemometrics Take Home 4 KU Leuven, Fall 2019

# Objective

The primary objective of this report is to demonstrate different clustering and classification methods. The data to be used will be of spectra gathered from grain samples using the Poly1720 camera. Four different grain samples were examined for a total of 674 calibration samples with 309 belonging to class 1, 64 belonging to class 2, 55 in class 3 and 246 for class 4. Below is a plot of the spectra where observations are colored by class.



# Clustering

To begin a simple baseline principle component analysis is to be performed. Examination of the generated biplots show the data can be separated somewhat well within a four-component model. Model statistics and four of the seemingly most relevant biplots are provided below. Points in the biplot are colored according to the recorded class. As Shown in the plots it is possible to get a rather strong level of class separation with these four components. Although it may appear that class 1 has some issues separating from other classes, it should be noted that, in accordance with the biplots, all three of the other classes do find separation in some way. This implies that if a forth dimensional plot were to be taken of PC1:4 the class 1 separation would be as evident as the separation of classes 2:4.

## Model

Principal Components Analysis Model

Developed 04-Dec-2019 15:12:34.899

Author: kebro@DESKTOP-B4UA9E4

X-block: 674 by 404 (kebro@DESKTOP-B4UA9E4@20191204T143712.88971446 m:20191204143712.892)

Included: [ 1-674 ] [ 1-404 ]

Preprocessing: Mean Center

Num. PCs: 4

Algorithm: SVD

Cross validation: venetian blinds w/ 10 splits and blind thickness = 1

RMSEC: 0.00316531

RMSECV: 0.00501423

## SSQ Table

Percent Variance Captured by PCA Model

Principal Eigenvalue % Variance % Variance

Component of Captured Captured

Number Cov(X) This PC Total

--------- ---------- ---------- ----------

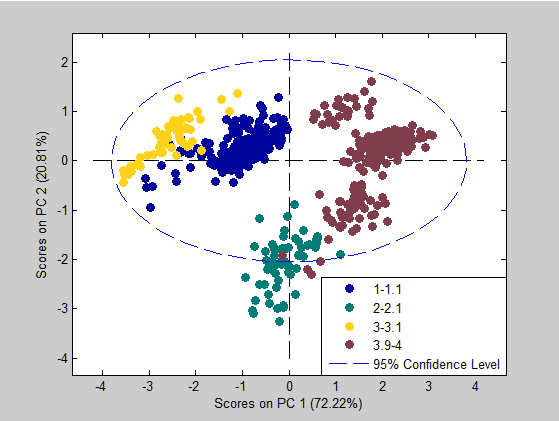
1 2.41e+00 72.22 72.22

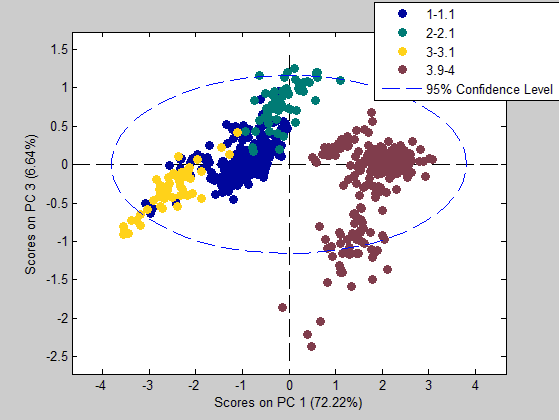
2 6.95e-01 20.81 93.03

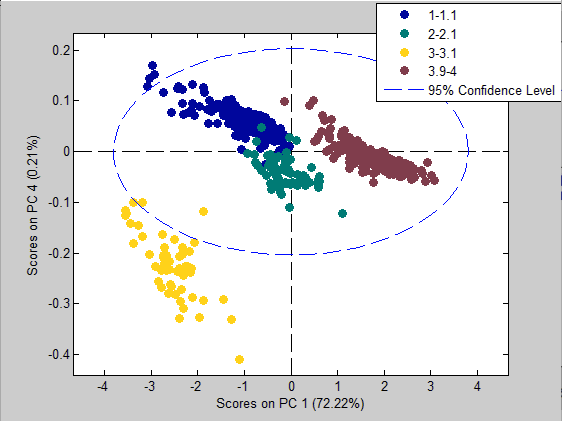
3 2.22e-01 6.64 99.67

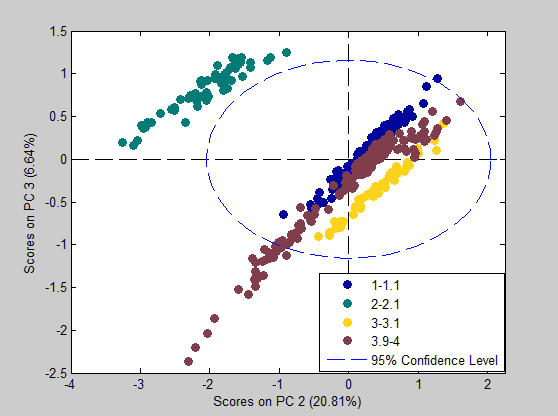
4 6.86e-03 0.21 99.88

## Figures associated with the analysis:









# Wards Method

Below are the resulting biplots of unsupervised clustering in accordance to Ward’s method. Principle components were chosen as inputs and the mahalanobis distance metric was used for the distance metric. As it can be clearly shown Ward’s method provides a near identical match to the baseline PCA clustering analysis. A notable remark is that class 1 does appear to have some overlapping with class 4 as seen in the comparison of the first quadrant of the baseline PCA and this method. This is a problem, but it appears to be unique with classes 1 and 4. A k-means model will be tested next to see if it can address this error.









# K-Means

Much like Ward’s method, class 1 and 4 are overlapping. A similar set up was used to the Ward’s method. Other classes are separated quite well, a further exploration using supervised methods should lead to stronger classification results.









# PLS-DA

Here a PLS-DA predictive model is trained using the grain spectra. To improve the model performance a first derivative was taken in addition to a mean centering. Doing so allowed for much better separation for class 1. To trim the model the data was filtered according to a genetic algorithm using a window width of 25. Results show a 5 latent variable model maps the data to their measured classes very well. Added more latent variables does nothing to address the cross-validation or calibration errors. There are very few outliers present in the data mostly belonging to class 2. This is odd as class two is one of the better predicted variables. Class prediction/probability charts for both the validation and calibration data are presented below along with relevant model statistics, calibration located on the left-hand side of the charts. From these visualizations, classes 1 and 3 are experiencing very minor overlap within the calibration data. This is not a problem for the prediction data where a seemingly perfect mapping is achieved.

## SSQ Table

Percent Variance Captured by Regression Model

-----X-Block----- -----Y-Block-----

Comp This Total This Total

---- ------- ------- ------- -------

1 74.98 74.98 28.32 28.32

2 17.53 92.51 32.41 60.73

3 5.05 97.55 23.89 84.62

4 1.61 99.16 7.96 92.58

5 0.29 99.45 3.06 95.64

## Prediction

This is a model of type: PLSDA\_PRED

Developed 05-Dec-2019 11:28:43.743

Author: kebro@DESKTOP-B4UA9E4

X-block: Grain\_Xtest 490 by 154 (kebro@DESKTOP-B4UA9E4@20191204T214438.47918590 m:20191204214627.580)

Included: [ 1-490 ] [ 1-25 51-100 176-200 226-250 276-300 401-404 ]

Preprocessing: 1st Derivative (order: 2, window: 15 pt, tails: polyinterp), Mean Center

Y-block: y 490 by 4 (kebro@DESKTOP-B4UA9E4@20191205T112843.66092337 m:20191205112843.660)

Included: [ 1-490 ] [ 1-4 ]

Preprocessing: Autoscale

Num. LVs: 5

Cross validation: random samples w/ 10 splits and 5 iterations

Statistics for each y-block column:

Sensitivity (Cal): 1.000 1.000 1.000 1.000

Specificity (Cal): 0.997 1.000 1.000 1.000

Sensitivity (CV): 1.000 1.000 1.000 1.000

Specificity (CV): 0.997 1.000 1.000 1.000

Sensitivity (Pred): 1.000 1.000 1.000 1.000

Specificity (Pred): 1.000 1.000 1.000 1.000

Class. Err (Cal): 0.00136986 0 0 0

Class. Err (CV): 0.00136986 0 0 0

Class. Err (Pred): 0 0 0 0

RMSEC: 0.131636 0.0221099 0.0492187 0.124408

RMSECV: 0.134328 0.0226136 0.0502193 0.126443

RMSEP: 0.130246 0.0202793 0.0489788 0.122012

Bias: 6.66134e-16 -1.66533e-16 -5.55112e-17 -5.55112e-16

CV Bias: 7.04052e-05 -2.61832e-05 2.16412e-05 -6.58633e-05

Pred Bias:-0.0134628 0.00071485 -0.00253545 0.0152835

R^2 Cal: 0.930206 0.994312 0.967676 0.933221

R^2 CV: 0.927472 0.994061 0.966407 0.93114

R^2 Pred: 0.928258 0.993271 0.954127 0.922175























# SVMDA

Now a support vector machine shall be used for the analysis. With similar results to the PLS-DA model there appears to be a perfect mapping of the cross-validation data. This can be seen in the confusion matrices where the prediction table shows two class 4 observations being mipmapped as class 1. Aside from that, the model performs very well. In this model, 98 support vectors were used to build the necessary discrimination boundaries. Observed error rates are approximately 0, with other important model statistics and relevant plots found below.

## Prediction

This is a model of type: SVMDA\_PRED

Developed 05-Dec-2019 12:13:24.506

Author: kebro@DESKTOP-B4UA9E4

X-block: Grain\_Xtest 490 by 154 (kebro@DESKTOP-B4UA9E4@20191204T214438.47918590 m:20191204214627.580)

Included: [ 1-490 ] [ 1-25 51-100 176-200 226-250 276-300 401-404 ]

Preprocessing: 1st Derivative (order: 2, window: 15 pt, tails: polyinterp), Mean Center

Y-block: 490 by 1 ( m:)

Included: [ 1-490 ] [ 1 ]

Preprocessing: None

SVM type: C-SVC

SVM kernel type: radial basis function

SVM optimal parameters:

cost = 100

gamma = 10

SVM: number of SVs: 98

Cross validation: custom (user) split

Statistics for each y-block column:

Sensitivity (Cal): 1.000 1.000 1.000 1.000

Specificity (Cal): 1.000 1.000 1.000 1.000

Sensitivity (CV): 1.000 1.000 1.000 0.967

Specificity (CV): 0.978 1.000 1.000 1.000

Sensitivity (Pred): 0.994 1.000 1.000 1.000

Specificity (Pred): 1.000 1.000 1.000 0.995

Class. Err (Cal): 0.00547945 0 0 0.00813008

Class. Err (CV): 0.0109589 0 0 0.0162602

Class. Err (Pred): 0.00324675 0 0 0.0027248

SVMDA Classification Using Rule: Pred Most Probable

MODEL RESULTS

Confusion Matrix:

Class: TPR FPR TNR FNR N Err P F1

1 1.00000 0.00000 1.00000 0.00000 309 0.00000 1.00000 1.00000

2 1.00000 0.00000 1.00000 0.00000 64 0.00000 1.00000 1.00000

3 1.00000 0.00000 1.00000 0.00000 55 0.00000 1.00000 1.00000

4 1.00000 0.00000 1.00000 0.00000 246 0.00000 1.00000 1.00000

Confusion Table:

Actual Class

1 2 3 4

Predicted as 1 309 0 0 0

Predicted as 2 0 64 0 0

Predicted as 3 0 0 55 0

Predicted as 4 0 0 0 246

Predicted as Unassigned 0 0 0 0

CV RESULTS

Confusion Matrix (CV):

Class: TPR FPR TNR FNR N Err P F1

1 1.00000 0.02192 0.97808 0.00000 309 0.01187 0.97476 0.98722

2 1.00000 0.00000 1.00000 0.00000 64 0.00000 1.00000 1.00000

3 1.00000 0.00000 1.00000 0.00000 55 0.00000 1.00000 1.00000

4 0.96748 0.00000 1.00000 0.03252 246 0.01187 1.00000 0.98347

Confusion Table (CV):

Actual Class

1 2 3 4

Predicted as 1 309 0 0 8

Predicted as 2 0 64 0 0

Predicted as 3 0 0 55 0

Predicted as 4 0 0 0 238

Predicted as Unassigned 0 0 0 0

PREDICTION RESULTS

Confusion Matrix:

Class: TPR FPR TNR FNR N Err P F1

1 0.99351 0.00000 1.00000 0.00649 308 0.00408 1.00000 0.99674

2 1.00000 0.00000 1.00000 0.00000 32 0.00000 1.00000 1.00000

3 1.00000 0.00000 1.00000 0.00000 27 0.00000 1.00000 1.00000

4 1.00000 0.00545 0.99455 0.00000 123 0.00408 0.98400 0.99194

Confusion Table:

Actual Class

1 2 3 4

Predicted as 1 306 0 0 0

Predicted as 2 0 32 0 0

Predicted as 3 0 0 27 0

Predicted as 4 2 0 0 123

Predicted as Unassigned 0 0 0 0

------------------------------------------------------------------------------------------------------

SVMDA Classification Using Rule: Pred Strict (using strictthreshold = 0.50)

MODEL RESULTS

Confusion Matrix:

Class: TPR FPR TNR FNR N Err P F1

1 1.00000 0.00000 1.00000 0.00000 309 0.00000 1.00000 1.00000

2 1.00000 0.00000 1.00000 0.00000 64 0.00000 1.00000 1.00000

3 1.00000 0.00000 1.00000 0.00000 55 0.00000 1.00000 1.00000

4 1.00000 0.00000 1.00000 0.00000 246 0.00000 1.00000 1.00000

Confusion Table:

Actual Class

1 2 3 4

Predicted as 1 309 0 0 0

Predicted as 2 0 64 0 0

Predicted as 3 0 0 55 0

Predicted as 4 0 0 0 246

Predicted as Unassigned 0 0 0 0

CV RESULTS

Confusion Matrix (CV):

Class: TPR FPR TNR FNR N Err P F1

1 1.00000 0.02192 0.97808 0.00000 309 0.01187 0.97476 0.98722

2 1.00000 0.00000 1.00000 0.00000 64 0.00000 1.00000 1.00000

3 1.00000 0.00000 1.00000 0.00000 55 0.00000 1.00000 1.00000

4 0.96748 0.00000 1.00000 0.03252 246 0.01187 1.00000 0.98347

Confusion Table (CV):

Actual Class

1 2 3 4

Predicted as 1 309 0 0 8

Predicted as 2 0 64 0 0

Predicted as 3 0 0 55 0

Predicted as 4 0 0 0 238

Predicted as Unassigned 0 0 0 0

PREDICTION RESULTS

Confusion Matrix:

Class: TPR FPR TNR FNR N Err P F1

1 0.99351 0.00000 1.00000 0.00649 308 0.00408 1.00000 0.99674

2 1.00000 0.00000 1.00000 0.00000 32 0.00000 1.00000 1.00000

3 1.00000 0.00000 1.00000 0.00000 27 0.00000 1.00000 1.00000

4 1.00000 0.00545 0.99455 0.00000 123 0.00408 0.98400 0.99194

Confusion Table:

Actual Class

1 2 3 4

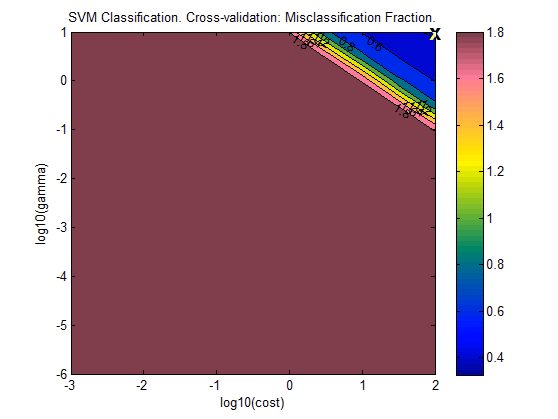
Predicted as 1 306 0 0 0

Predicted as 2 0 32 0 0

Predicted as 3 0 0 27 0

Predicted as 4 2 0 0 123

Predicted as Unassigned 0 0 0 0



















# K Nearest Neighbors

To conclude the supervised methods, a KNN model is to be evaluated. From the table below it can be seen that a KNN model leads to perfect classification. A model with 2, 3, 10 and 20 neighbors were created. All lead to the same result. Below are the model statistics for the 20-neighbor model. An error rate of 0 was calculated for both the cross-validation and prediction data. Found below are relevant model measurements and plots. From the plots, it can be discerned that the model struggles most with separating classes 1 and 4. This is unsurprising as it is in line with the previous model’s performance. Fortunately, the KNN model does overcome the problems associated with the past models.

## Prediction

This is a model of type: KNN\_PRED

Developed 05-Dec-2019 22:10:11.313

Author: kebro@DESKTOP-B4UA9E4

X-block: Grain\_Xtest 490 by 154 (kebro@DESKTOP-B4UA9E4@20191204T214438.47918590 m:20191205221010.476)

Included: [ 1-490 ] [ 1-25 51-100 176-200 226-250 276-300 401-404 ]

Preprocessing: 1st Derivative (order: 2, window: 15 pt, tails: polyinterp), Mean Center

Cross validation: random samples w/ 10 splits and 5 iterations

Statistics for each y-block column:

Sensitivity (Cal): 1.000 1.000 1.000 1.000

Specificity (Cal): 1.000 1.000 1.000 1.000

Sensitivity (CV): 1.000 1.000 1.000 1.000

Specificity (CV): 1.000 1.000 1.000 1.000

Sensitivity (Pred): 1.000 1.000 1.000 1.000

Specificity (Pred): 1.000 1.000 1.000 1.000

Class. Err (Cal): 0 0 0 0

Class. Err (CV): 0 0 0 0

Class. Err (Pred): 0 0 0 0

KNN Classification Using Rule: Pred Most Probable

MODEL RESULTS

Confusion Matrix:

Class: TPR FPR TNR FNR N Err P F1

1 1.00000 0.00000 1.00000 0.00000 309 0.00000 1.00000 1.00000

2 1.00000 0.00000 1.00000 0.00000 64 0.00000 1.00000 1.00000

3 1.00000 0.00000 1.00000 0.00000 55 0.00000 1.00000 1.00000

4 1.00000 0.00000 1.00000 0.00000 246 0.00000 1.00000 1.00000

Confusion Table:

Actual Class

1 2 3 4

Predicted as 1 309 0 0 0

Predicted as 2 0 64 0 0

Predicted as 3 0 0 55 0

Predicted as 4 0 0 0 246

Predicted as Unassigned 0 0 0 0

CV RESULTS

Confusion Matrix (CV):

Class: TPR FPR TNR FNR N Err P F1

1 1.00000 0.00000 1.00000 0.00000 309 0.00000 1.00000 1.00000

2 1.00000 0.00000 1.00000 0.00000 64 0.00000 1.00000 1.00000

3 1.00000 0.00000 1.00000 0.00000 55 0.00000 1.00000 1.00000

4 1.00000 0.00000 1.00000 0.00000 246 0.00000 1.00000 1.00000

Confusion Table (CV):

Actual Class

1 2 3 4

Predicted as 1 309 0 0 0

Predicted as 2 0 64 0 0

Predicted as 3 0 0 55 0

Predicted as 4 0 0 0 246

Predicted as Unassigned 0 0 0 0

PREDICTION RESULTS

Confusion Matrix:

Class: TPR FPR TNR FNR N Err P F1

1 1.00000 0.00000 1.00000 0.00000 308 0.00000 1.00000 1.00000

2 1.00000 0.00000 1.00000 0.00000 32 0.00000 1.00000 1.00000

3 1.00000 0.00000 1.00000 0.00000 27 0.00000 1.00000 1.00000

4 1.00000 0.00000 1.00000 0.00000 123 0.00000 1.00000 1.00000

Confusion Table:

Actual Class

1 2 3 4

Predicted as 1 308 0 0 0

Predicted as 2 0 32 0 0

Predicted as 3 0 0 27 0

Predicted as 4 0 0 0 123

Predicted as Unassigned 0 0 0 0

------------------------------------------------------------------------------------------------------

KNN Classification Using Rule: Pred Strict (using strictthreshold = 0.50)

MODEL RESULTS

Confusion Matrix:

Class: TPR FPR TNR FNR N Err P F1

1 1.00000 0.00000 1.00000 0.00000 309 0.00000 1.00000 1.00000

2 1.00000 0.00000 1.00000 0.00000 64 0.00000 1.00000 1.00000

3 1.00000 0.00000 1.00000 0.00000 55 0.00000 1.00000 1.00000

4 1.00000 0.00000 1.00000 0.00000 246 0.00000 1.00000 1.00000

Confusion Table:

Actual Class

1 2 3 4

Predicted as 1 309 0 0 0

Predicted as 2 0 64 0 0

Predicted as 3 0 0 55 0

Predicted as 4 0 0 0 246

Predicted as Unassigned 0 0 0 0

CV RESULTS

Confusion Matrix (CV):

Class: TPR FPR TNR FNR N Err P F1

1 1.00000 0.00000 1.00000 0.00000 309 0.00000 1.00000 1.00000

2 1.00000 0.00000 1.00000 0.00000 64 0.00000 1.00000 1.00000

3 1.00000 0.00000 1.00000 0.00000 55 0.00000 1.00000 1.00000

4 1.00000 0.00000 1.00000 0.00000 246 0.00000 1.00000 1.00000

Confusion Table (CV):

Actual Class

1 2 3 4

Predicted as 1 309 0 0 0

Predicted as 2 0 64 0 0

Predicted as 3 0 0 55 0

Predicted as 4 0 0 0 246

Predicted as Unassigned 0 0 0 0

PREDICTION RESULTS

Confusion Matrix:

Class: TPR FPR TNR FNR N Err P F1

1 1.00000 0.00000 1.00000 0.00000 308 0.00000 1.00000 1.00000

2 1.00000 0.00000 1.00000 0.00000 32 0.00000 1.00000 1.00000

3 1.00000 0.00000 1.00000 0.00000 27 0.00000 1.00000 1.00000

4 1.00000 0.00000 1.00000 0.00000 123 0.00000 1.00000 1.00000

Confusion Table:

Actual Class

1 2 3 4

Predicted as 1 308 0 0 0

Predicted as 2 0 32 0 0

Predicted as 3 0 0 27 0

Predicted as 4 0 0 0 123

Predicted as Unassigned 0 0 0 0

# Conclusion

Clustering methods provide a way to separate observations into classes which may not be initially apparent. Here unsupervised and supervised methods were cross-compared and determined to be quite strong. The supervised methods do prove to be quite powerful, providing perfect or near perfect mappings. Class 1 and class 4 appear to be quite similar so their classifications are typically less than ideal but still rather strong despite this. Overall, the methods used in this report prove to provide workable classification results for this particular data set.