Treatment of Environmental Data
Questions and Answers

Part 1

Correlation Table				
	sucrose	*glucose*	*fructose*	*sorbitol*
sucrose	1			
glucose	-0.77474	1		
fructose	-0.38047	0.373303	1	
sorbitol	-0.06675	0.232694	0.186765	1

1. Do any of the columns appear to be highly correlated?

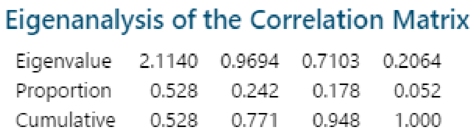
Only sucrose and glucose have a somewhat strong albeit negative correlation at -0.77. All other sugars do not show any significant relationship.

PCA analysis (Minitab)

2. What is the appropriate pre-treatment?

The appropriate pre-treatment is the mean-centre because the data variables are parameters are standardized and measure the concentration of different sugars within apple juice in g/L. We don't need to standardize it because it's not on different scales.

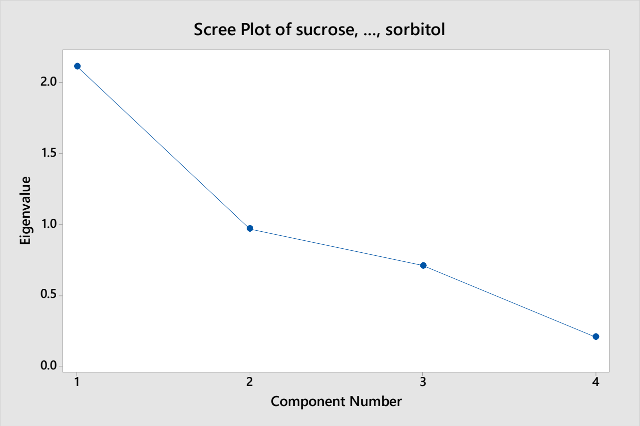
3. How many latent variables appear sufficient to describe most of the variability in the data? How much is explained by the first 2 principal components?



Eigenanalysis of the Correlation Matrix

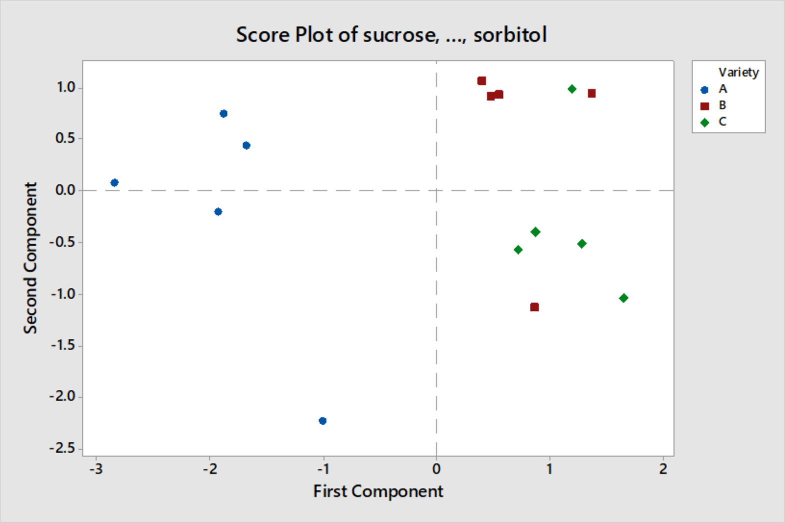
Eigenvalue	2.1140	0.9694	0.7103	0.2064
Proportion	0.528	0.242	0.178	0.052
Cumulative	0.528	0.771	0.948	1.000

53% explains the total variability of the data. The second principal component has 0.24 and accounts for 77% of the variance in the data. Most of the data is explained by the first two components.



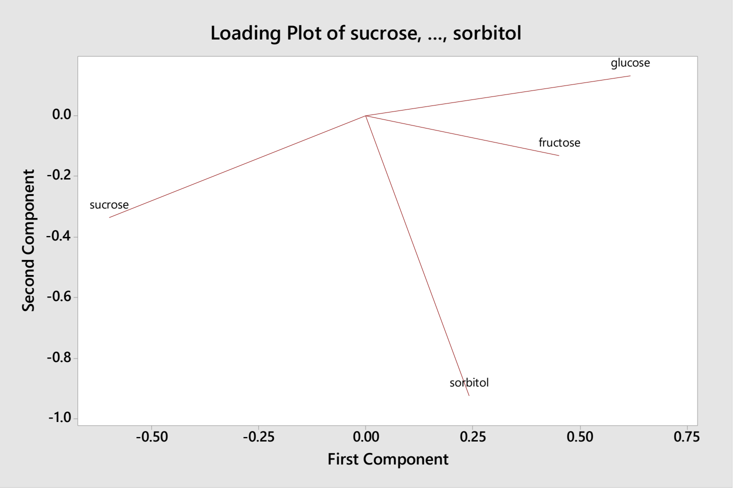
The scree plot shows that the most significant contributions are the first and second principal components that fall dramatically then steadily with component 3 and 4.

4. Produce a score plot of the first 2 principal components. Is there any apparent grouping? Does it appear to coincide with the fresh and stored groups? If there is separation along either axis look at the loadings and comment on which ones most affect this separation



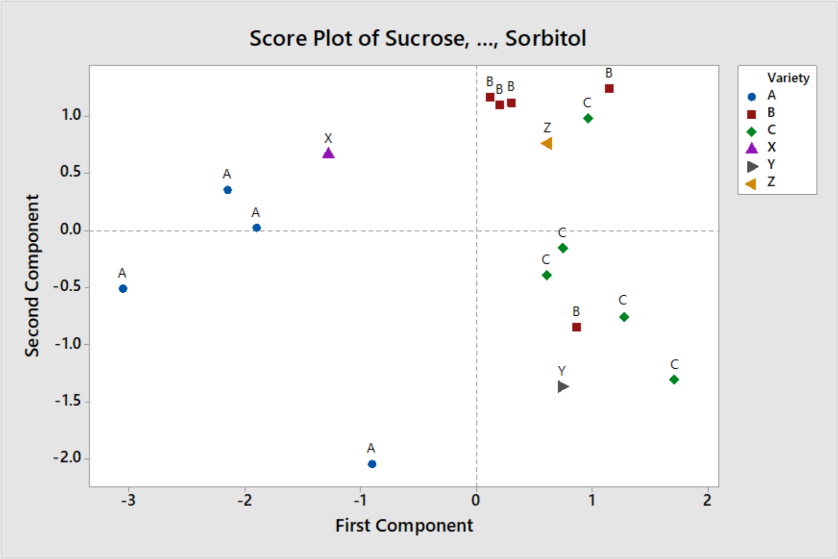
There are clear groupings where A is on the left side separated from both B and C which are on the right. It has been effective at separating A from B and C. However, majority of B are on the top right quadrant whereas C is on the bottom right quadrant with a single outlier from each group.

5. Do a loadings plot of the first 2 variables. Do any variables appear to be associated?



The loading plot visually shows the results for the first two components. Sucrose has a negative loading on the second and first component. Sorbitol has a positive loading for the first component. It is fructose and glucose that appear to be associated together as they lean with a positive loading for the first component. Majority of the variables have a negative loading on the second component.

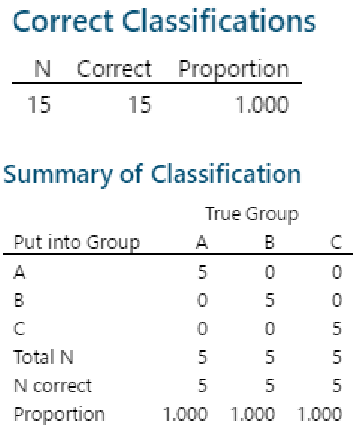
PCA and Unknowns



6. Comment overall on the usefulness of PCA with this data set

PCA was extremely useful with the sugars in apple juice data set. Variety A is separated from variety B and C as it is on the other side of the origin plot. It shows that variety A is negatively correlated and the variables in A have more impact from both the second and first components as they are spread further away from the origin. Variety B are clustered together in the upper right quadrant closest to the origin line and has one variety C outlier grouped together. Variety C are grouped together in the lower right quadrant and has a single outlier from variety B. All the unknowns were successfully allocated with their respective varieties; as X belongs with variety A, Y belongs with variety C and Z belongs with variety B.

LDA Analysis



Correct Classifications

N	Correct	Proportion
15	15	1.000

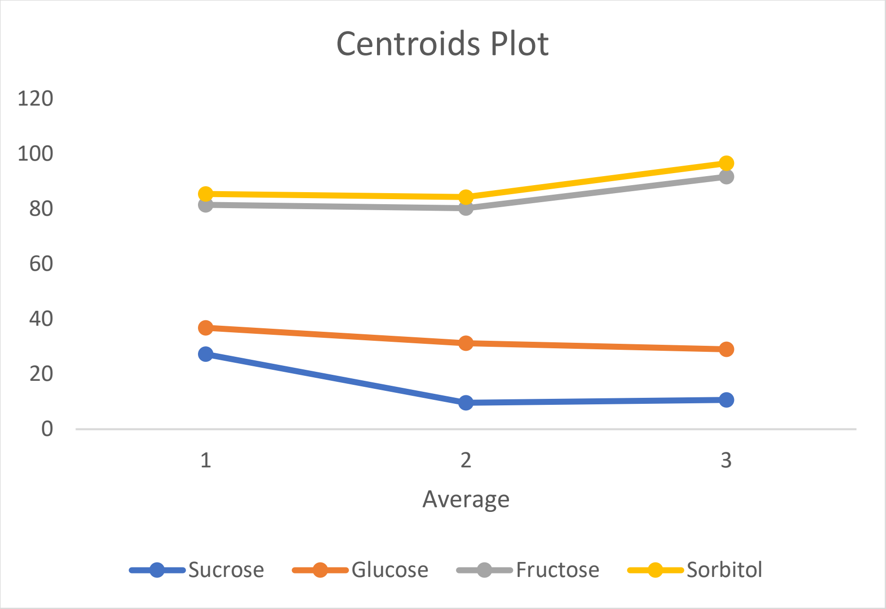
Summary of Classification

	True Group		
Put into Group	A	B	C
A	5	0	0
B	0	5	0
C	0	0	5
Total N	5	5	5
N correct	5	5	5
Proportion	1.000	1.000	1.000

7. From the misclassification plot how successful is LDA in assigning groups?

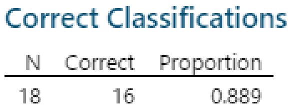
LDA was perfect because all 15 datapoints were correctly assigned at 100%.

8. Do any the variables appear to be better at discriminating groups (examine the centroids plot)



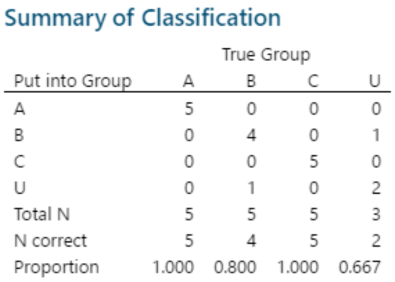
According to the centroids plot, sorbitol and fructose appear to be highly correlated and grouped together. Sucrose and glucose are similar and found together. The wide gap indicates that there are two distinct similar groups for the 4 sugars.

9. Use LDA to predict the groups for the unknowns



Correct Classifications

N	Correct	Proportion
18	16	0.889



Summary of Classification

	True Group			
Put into Group	A	B	C	U
A	5	0	0	0
B	0	4	0	1
C	0	0	5	0
U	0	1	0	2
Total N	5	5	5	3
N correct	5	4	5	2
Proportion	1.000	0.800	1.000	0.667

The LDA was somewhat successful at assigning correct groups overall as 16 out of 18 were assigned correctly at 89%. It is assigning the unknowns that has 2 out of 3 were correct but a low 67% was assigned correctly.

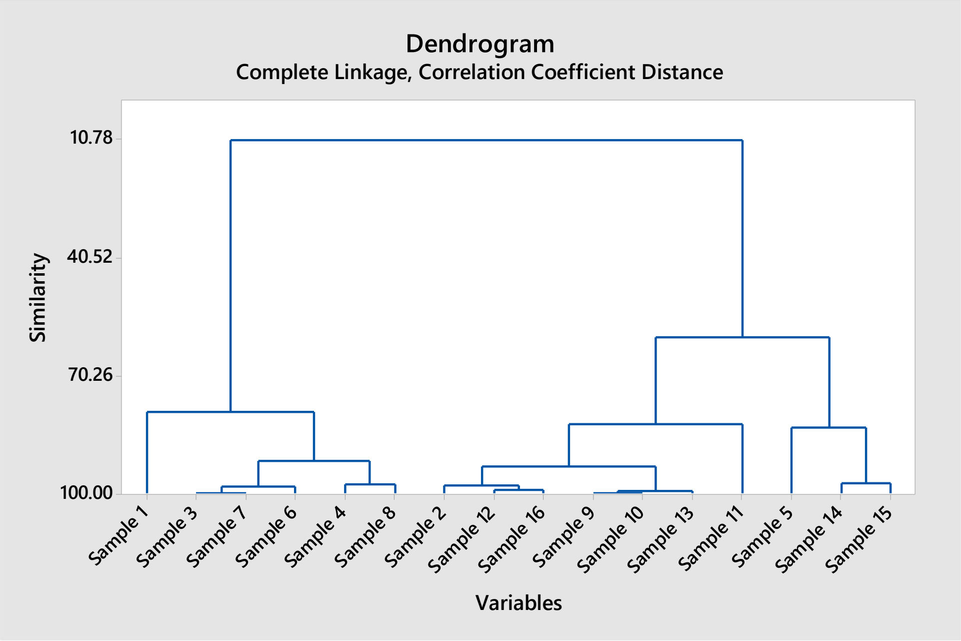
10. Compare PCA and LDA for predicting group membership

PCA produced a score plot which visually showed how unknown sources can be grouped with the three apple juices varieties. It made predictions based on correlations and maximized separation based on variance. PCA is an unsupervised data reduction method. Whereas LDA is based on mean vectors and assumes that the data is normally distributed. It was useful at providing a number quantification of predicting group membership, but it does not provide visual data.

Part 2

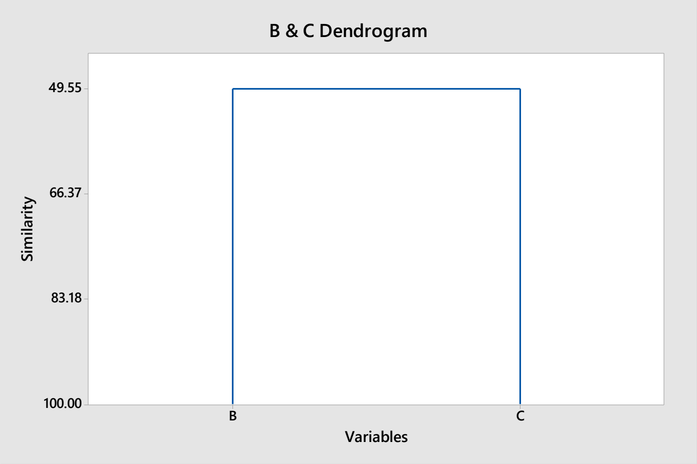
Hierarchical Analysis

1. Create a dendrogram for the above data set, using Minitab and all 5 variables. Do you see any groupings and do they agree with the groupings above.



There are two obvious groupings. Sample 1, 3, 7, 6, 4 & 8 appear to be highly correlated whereas sample 2, 12, 16, 9, 10, 13, 11, 5, 14 & 15 are grouped together.

2. Repeat with just the 2 columns assigned to you. Do you get the same groupings?



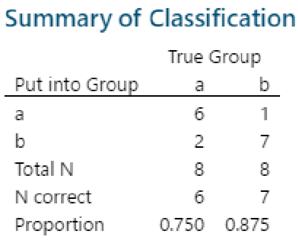
The grouping for B & C produce very little information and simply show a 50% similarity.

NMC Analysis

As attached on the spreadsheet.

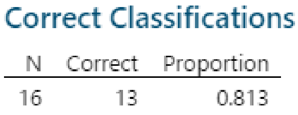
LDA

3. Do LDA using just your 2 assigned columns, including prediction of the unknowns.



Summary of Classification

Put into Group	True Group a	 b
a	6	1
b	2	7
Total N	8	8
N correct	6	7
Proportion	0.750	0.875

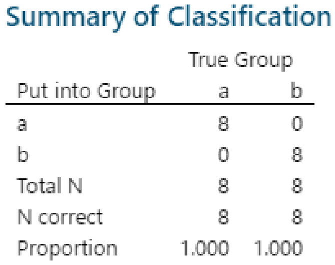


Correct Classifications

N	Correct	Proportion
16	13	0.813

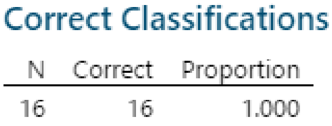
The LDA method showed that 13 out of 16 were classified correctly at a total proportion of 81%. Group b had a higher number of correct groupings 7 out of 8 at 88% whereas group a only had 6 out of 8 correct at 75%.

4. Repeat using all 5 columns.



Summary of Classification

Put into Group	True Group a	 b
a	8	0
b	0	8
Total N	8	8
N correct	8	8
Proportion	1.000	1.000



Correct Classifications

N	Correct	Proportion
16	16	1.000

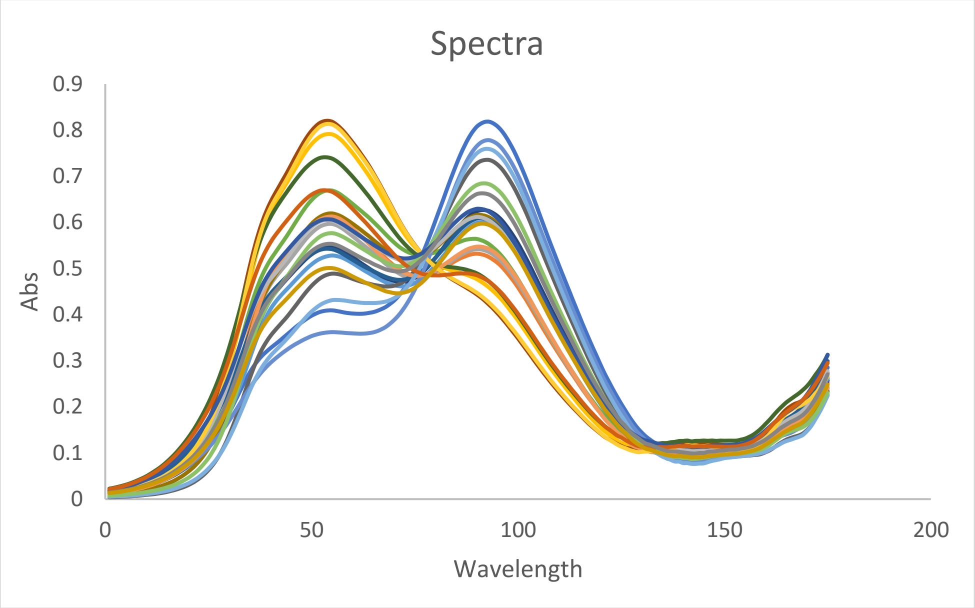
When all 5 columns are analysed, it shows that all variables have been classed correctly at a total proportion of 100%.

5. Comment on how each performs and how this result compares to the nearest means analysis.

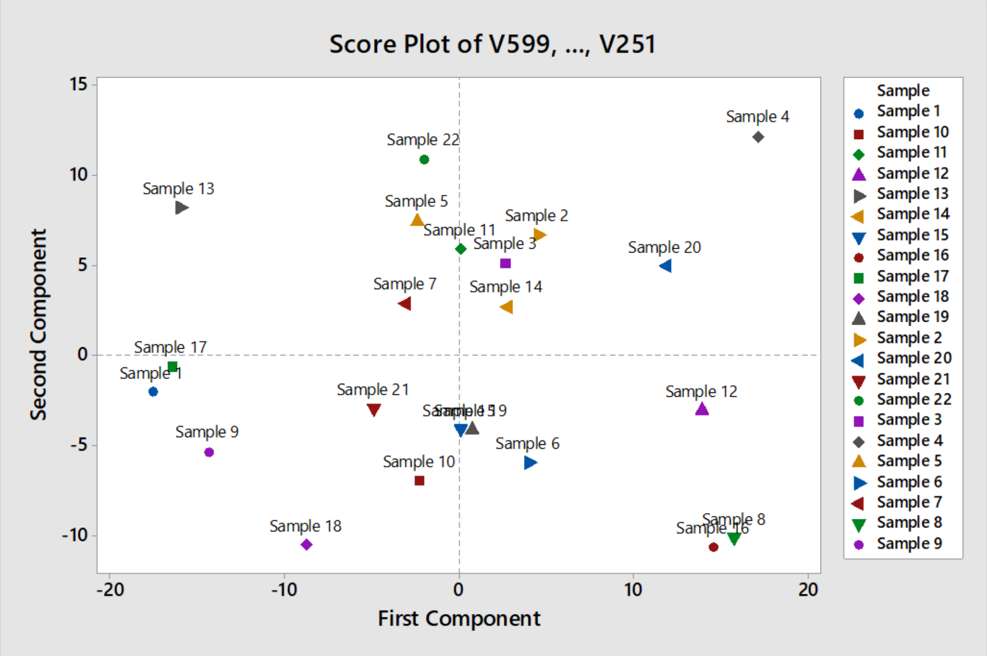
The NMC analysis had 4 misclassified, so for 12 out of a total of 16 for columns B and C was correctly allocated. As shown on the spreadsheet, sample 12 can be placed in both a/b which is almost identical to the LDA analysis that showed 13 out of 16 had been correctly grouped. Since NMC assumes the data is on the same scale according to the Euclidian distance and the LDA measures the Mahalanobis distance, they both result in very similar results.

Part 3

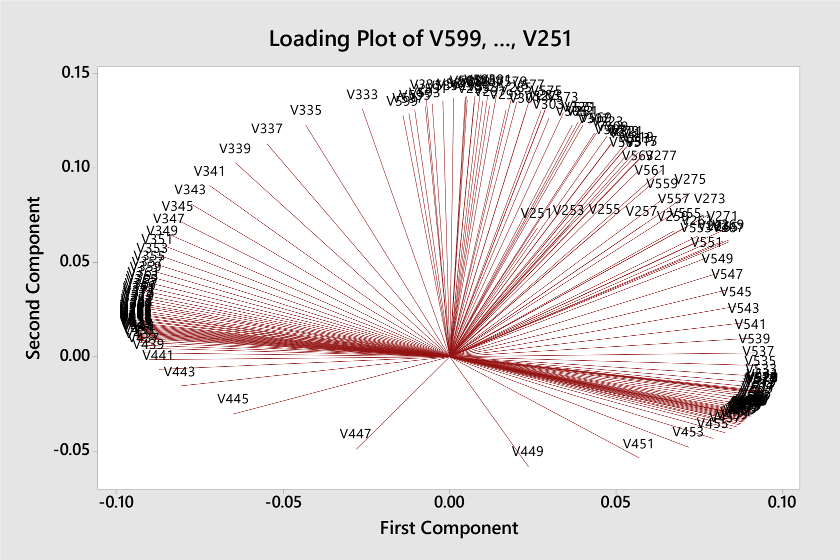
6. Plot on one graph all the spectra to see what the data looks like.



7. Carry out PCA on the whole absorption data set, using the appropriate pre-treatment. is there any evidence of outliers from the score plot?

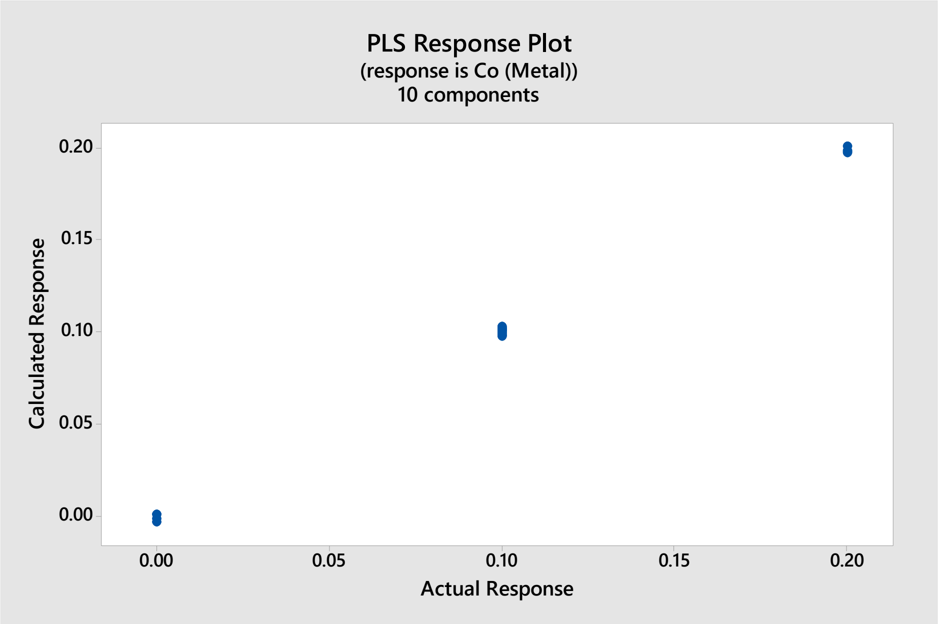


All 22 samples have an even spread across the score plot that indicate no obvious outliers. The samples were pre-treated with mean centring as PCA does not have any structure in the data.



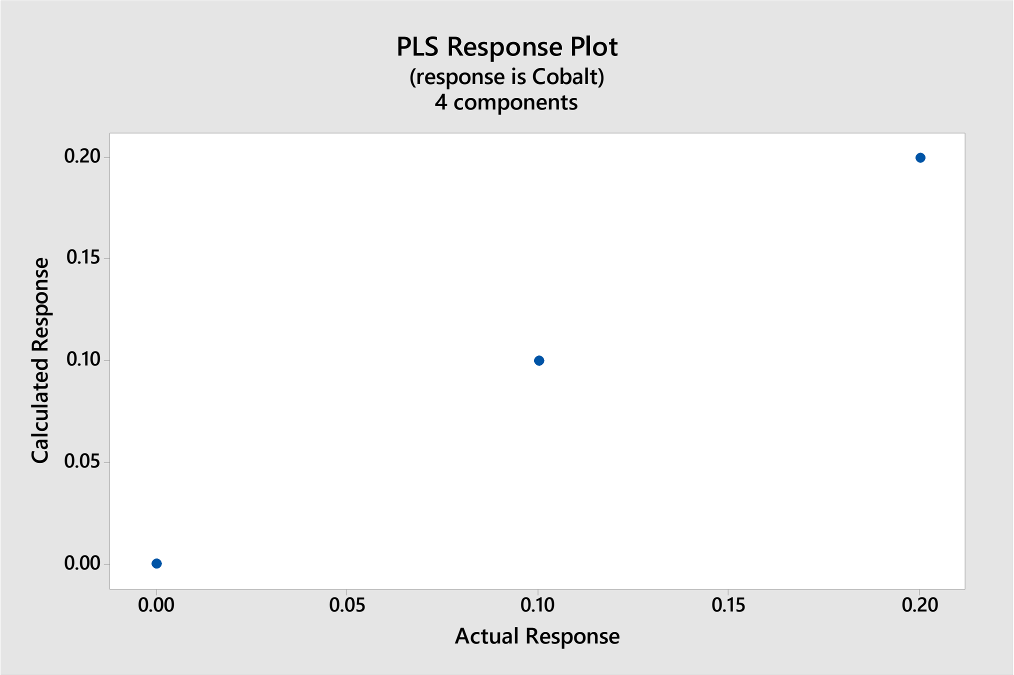
The loadings plot reaffirms the findings of the score plot as all wavelengths are evenly spread across.

8. Carry out PLS analysis using the wavelength data set and the concentrations of your assigned metal as responses. Use at least as many PCs as you have components (metals) and a few extra for noise.



The p-value was at 0.000 ($p<0.05$) which means that regression works very well for cobalt which can be seen on the PLS plot as actual vs. calculated response have produced a very positively linear organization of the datapoints.

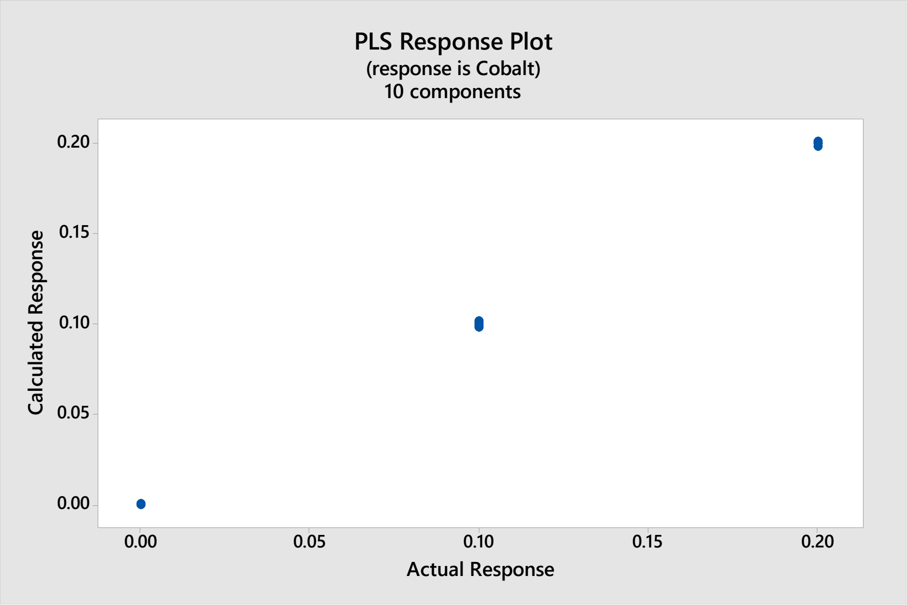
9. Decide on the optimal number of components (beware of overfitting) and plot a graph of predicted vs actual concentrations.



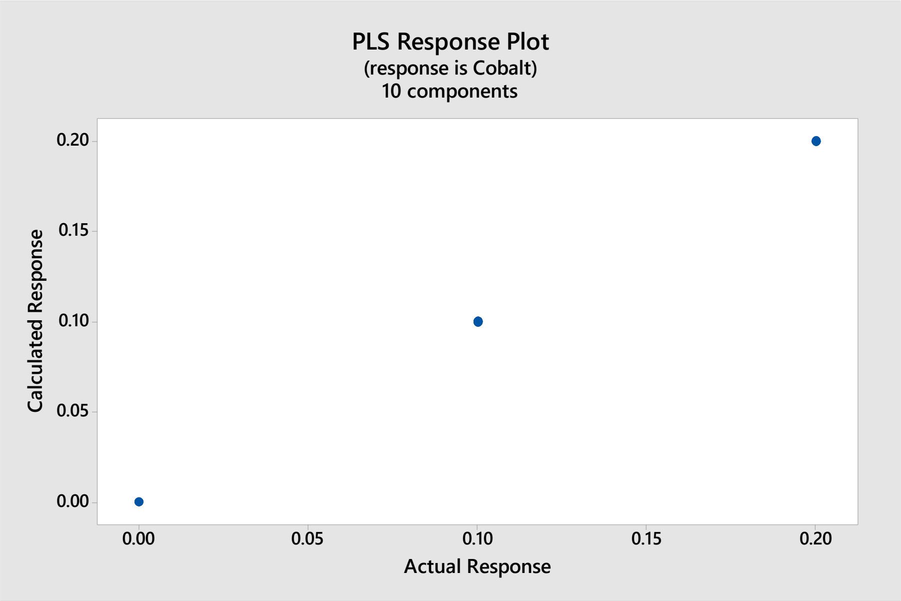
5 samples were chosen at random (sample 8, 10, 11, 14 & 18) to determine if running less samples produced a better fit. Cobalt produced a p-value of 0.000 which is identical to the regression carried out on all 22 samples. It shows that segmenting the dataset to show an optimal amount was not needed.

10. Calculate the first and second derivatives of each row using S-G polynomials (see sample sheet) and repeat PLS for each. How do they compare? Is there a marked improvement? Compare the performances using appropriate diagnostics.

First Derivative

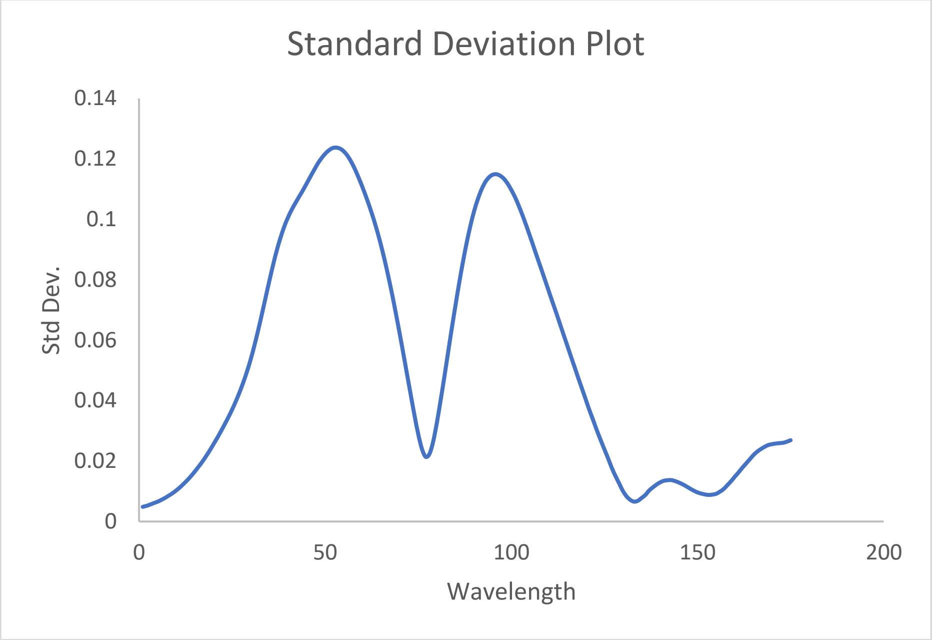


Second Derivative



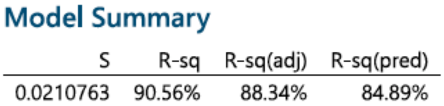
The first and second derivatives produced identical p-values at 0.000 ($p<0.05$) and response plots where the data moves in a positive linear direction. Cobalt works well without doing derivates and moving to the 2nd derivative was unnecessary. It produces a good fit with the original concentrations with cobalt.

11. Do a standard deviation plot of wavelength vs s.d. of each column



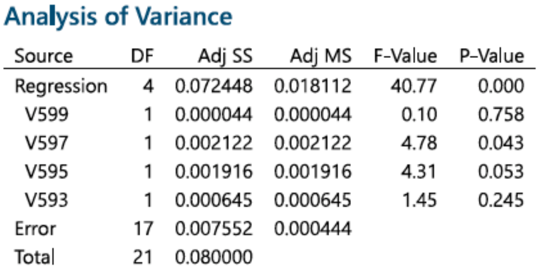
12. From this plot choose a subset of wavelengths (up to the number of samples) for ordinary least squares and do a multivariate regression for your metal. How does the result compare to PLS?

OLS (V599, V597, V595, V593)



Model Summary

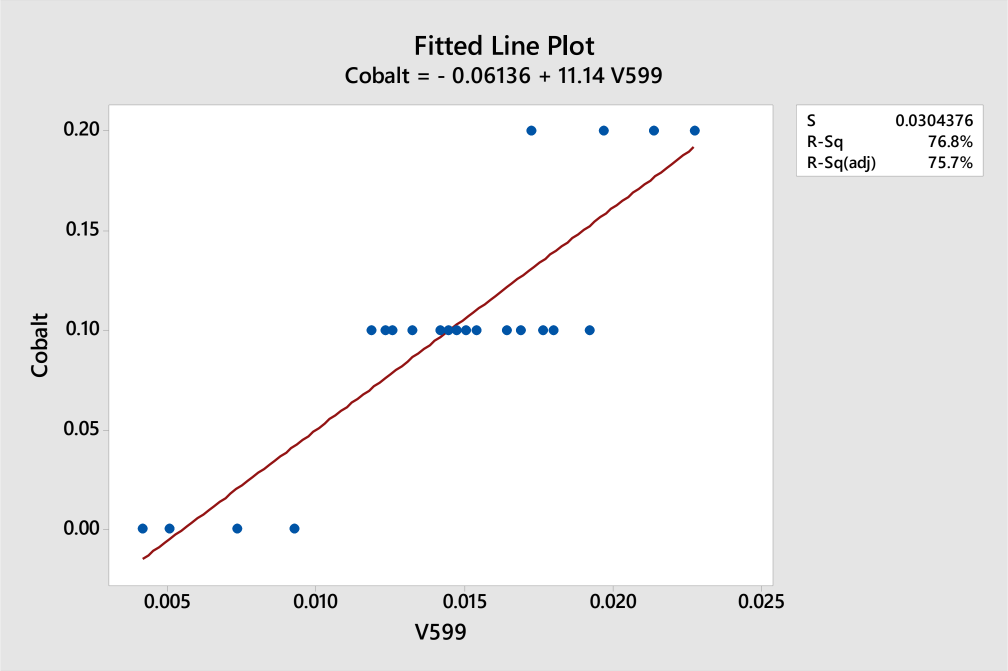
S	R-sq	R-sq(adj)	R-sq(pred)
0.0210763	90.56%	88.34%	84.89%



Analysis of Variance

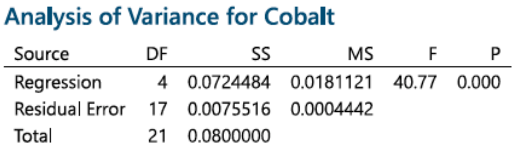
Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	4	0.072448	0.018112	40.77	0.000
V599	1	0.000044	0.000044	0.10	0.758
V597	1	0.002122	0.002122	4.78	0.043
V595	1	0.001916	0.001916	4.31	0.053
V593	1	0.000645	0.000645	1.45	0.245
Error	17	0.007552	0.000444		
Total	21	0.080000			

The multivariate regression for cobalt vs. a small subset of wavelengths (v599, v597, v595 & v593) show that the standard error is small at 0.02 and the data is highly statistically significant as the p-value is at 0.000 (p<0.05) and has a strong relationship variation at 90.6% (r-squared value).



The fitted line plot for the single wavelength V599 has a r-squared value of 76.8%. It is lower in comparison than the four subset of wavelengths.

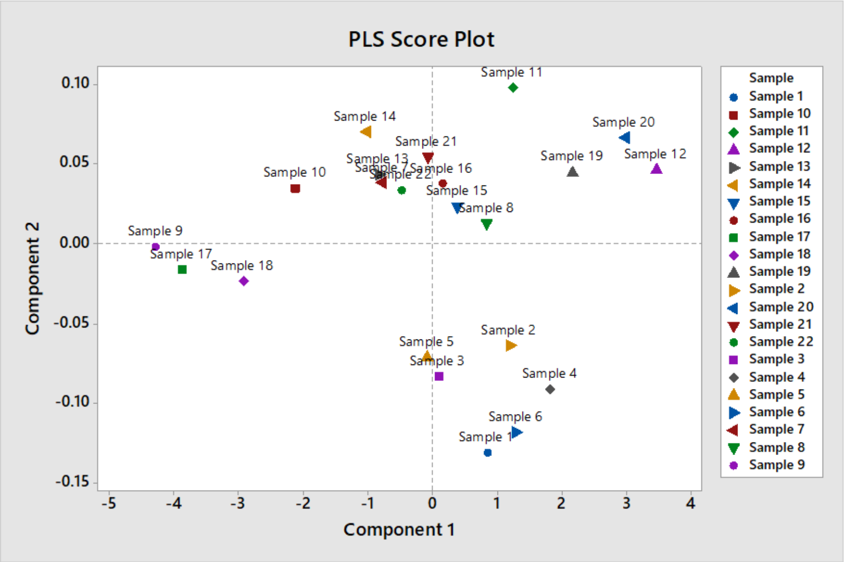
Partial Least Squares



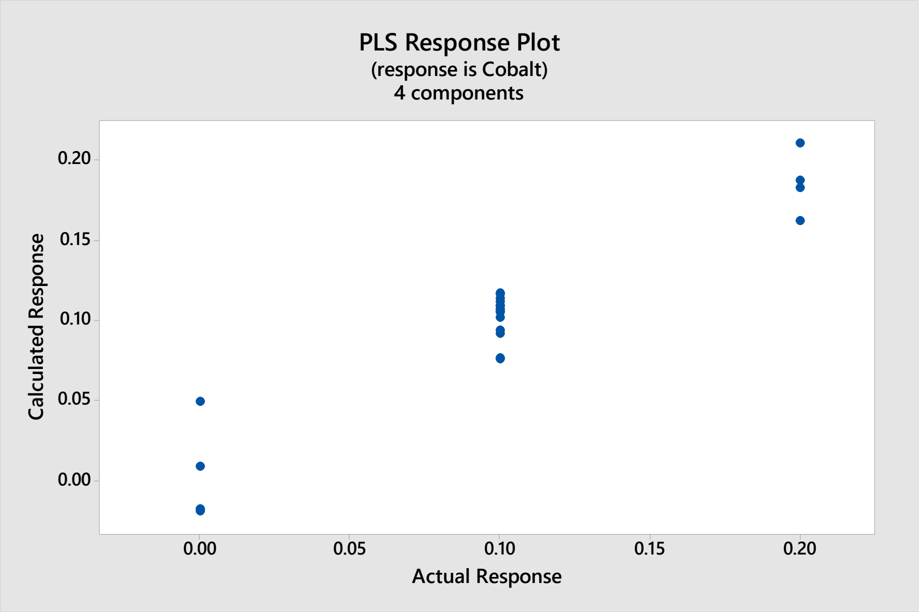
Analysis of Variance for Cobalt

Source	DF	SS	MS	F	P
Regression	4	0.0724484	0.0181121	40.77	0.000
Residual Error	17	0.0075516	0.0004442		
Total	21	0.0800000			

The PLS regression for cobalt uses the same small subset of wavelengths (v599, v597, v595 & v593). It produced the lowest residual error at 0.007 and like multivariate regression, the p-value is at 0.000 ($p < 0.05$). The r-squared value using PLS between cobalt vs. 599 results in a slightly higher variation of data at 79% than of the OLS result at 76.8%.



The score plot using four components shows grouping of samples. For example, sample 5, 3, 2, 4, 6 & 1 are distinctively grouped together on the lower right quadrant whereas sample 9, 17 & 18 are in the middle left quadrant separated from the remaining samples that are varied across the upper region of the plot.



The PLS plot with four components shows a more detailed variation than the previous methods of using first and second derivatives for 10 components. There is a wider spread, but the data still moves in a linear direction thus showing a subset and less datapoints for cobalt is less perfect.