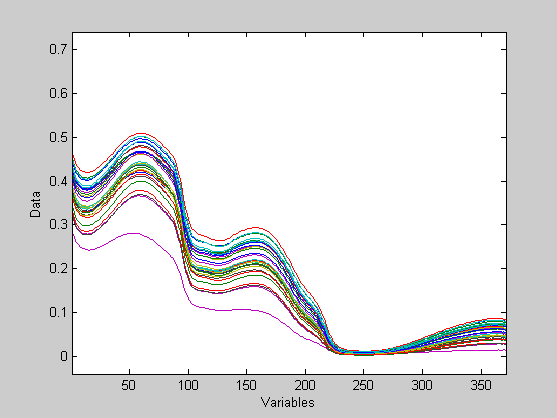
# Chemometrics Assignment 1 Kendall Brown, r0773111 KU Leuven, Session 1, 2019

An analysis is to be performed on data gathered from the spectral reflectance of 50 raw milk samples. The objective is to form a regression model which can reliably produce an accurate measurement of a given milk samples protein content.

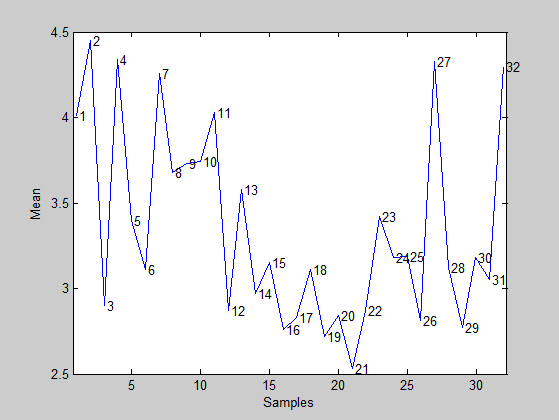
Regarding the dimensionality of the data. Given the dimensionality of our data set it would be impractical to create a regression model without first conducting a dimensionality reduction. We have over 350 spectra measurements for each of our 32 samples in the calibration set. This is an issue as without a dimensionality reduction we will likely run into problems regarding multi-collinearity and regression coefficient estimation. To address this problem, we will employ and compare the results of a principle components regression (PCR) model versus a partial least square regression (PLSR) model. These regression methods will compress our data to a size much more manageable for the purposes of building predictive models.

We shall begin by looking at the raw spectra of out calibration set. The spectra measurements were gathered from a Zeiss Corona Plus NIR 1.7 (950-1690 nm with 2 nm resolution).



Initially we see a strong downward trend indicating the presence of influential factors within our data set. These factors seem to be most influential within the 1070-1450nm range. When the data is to be used for regression, this range of values will be the primary area of focus for model training.

Now we shall analyze the data using principle component analysis. Found below will be the auto generated report of the PCA provided by SOLO (images replaced to show observation index). The raw data is to be mean centered prior to the analysis. A ten-fold, five iteration random subset cross-validation method was chosen as there does not appear to be a clear relation between milk samples (scatter plot of measured protein content showed below).



# Analysis Report

## Model

Principal Components Analysis Model

Developed 05-Nov-2019 13:23:11.487

Author: kebro@DESKTOP-B4UA9E4

X-block: Xcal 32 by 371 (kebro@DESKTOP-B4UA9E4@20191105T132208.42128183 m:20191105132208.421)

Included: [ 1-32 ] [ 1-371 ]

Preprocessing: Mean Center

Num. PCs: 8

Algorithm: SVD

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

RMSEC: 0.000103982

RMSECV: 0.000532373

## SSQ Table

Percent Variance Captured by PCA Model

Principal Eigenvalue % Variance % Variance

Component of Captured Captured

Number Cov(X) This PC Total

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

1 3.88e-01 98.74 98.74

2 4.06e-03 1.03 99.78

3 5.53e-04 0.14 99.92

4 2.18e-04 0.06 99.97

5 5.14e-05 0.01 99.99

6 3.01e-05 0.01 99.99

7 1.28e-05 0.00 100.00

8 6.17e-06 0.00 100.00

## Prediction

Principal Components Analysis Model

Developed 01-Nov-2019 21:21:44.157

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 371 (kebro@DESKTOP-B4UA9E4@20191101T184943.03351211 m:20191101184943.035)

Included: [ 1-15 ] [ 1-371 ]

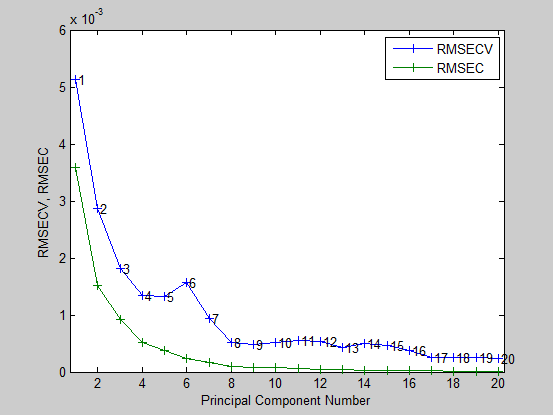
Preprocessing: Mean Center

Num. PCs: 8

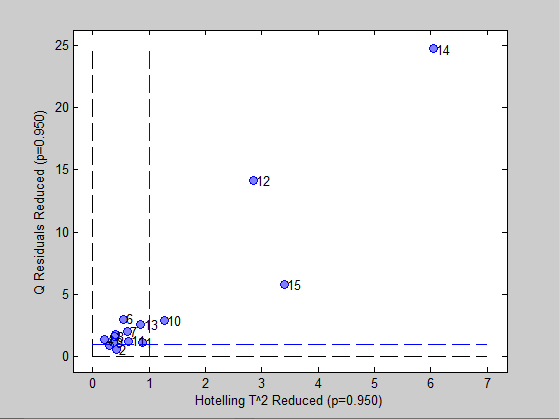
Algorithm: SVD

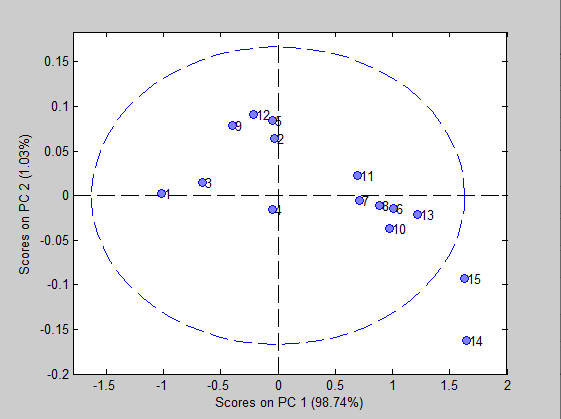
## Figures associated with the analysis:

## Model Statistics - PCA 8 PCs -

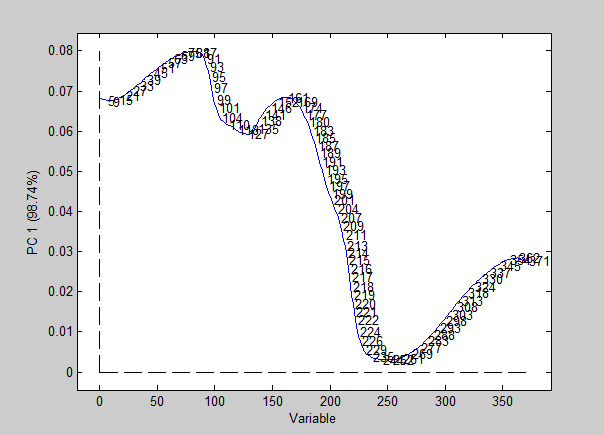


## Samples/Scores - PCA 8 PCs -

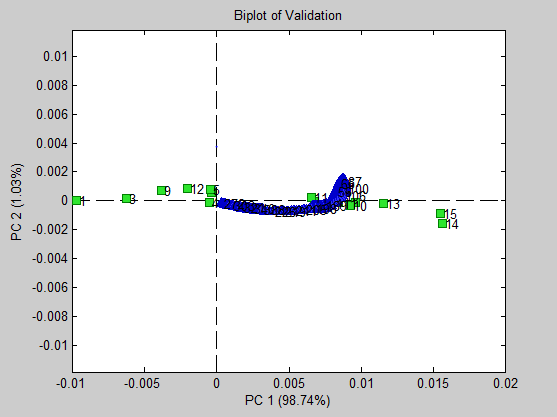




## Variables/Loadings - PCA 8 PCs -



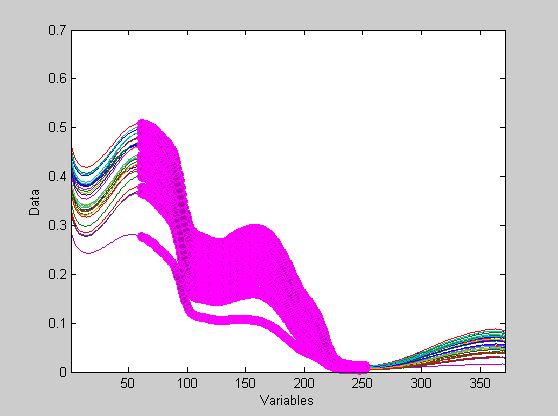
## Biplot - PCA 8 PCs -



As observed in the above plots and summary statistics, we see that nearly all the information can be described in within the first component (explains 98.74% of observed variance). Additionally, we see that an 8-component model describes the data quite well as the RMSECV tapers off after the addition of the eighth component. Two notable outliers, samples 14 and 15, appear in the biplots along with two distinct groupings. Upon comparison with the validation set’s protein measurements we could loosely described these groupings as being “high-protein” and “low-protein” samples. For our RMSE measurements we find our model to be more than adequate with a cross-validation RMSE of approximately .0013.

# Principle Component Regression

To begin our principle component regression, we first must select the most influential observations. As stated earlier, we see the largest shift in spectra from the highlighted region seen below (resolution 1070-1450nm).



Selecting the highlighted region to be the basis for which our model is calculated from we focus on the factors driving the change observed in spectra. Additionally, we see that observation 5 is injecting quite a bit of noise into the data set. It will be removed for the purposes of building the models. Outside of mean centering, no advanced pre-processing method will be used as per assignment instructions.

The results of the analysis show that a 13-component regression model carries adequate validation and prediction results. As seen in the plots below, including more than 13 components is seemingly pointless as it does not provide significant improvements. We calculated a RMSE of cross-validation to approximately 0.209, a R^2 of cross-validation of 0.93, and a cross-validation bias of -0.008. Fortunately, we see our model produces prediction results generally worse than those of our validation. We see a RMSE of prediction of 0.278, a R^2 prediction of 0.837 and a prediction bias of 0.0561. This implies that our model is not generalizing to new data as well as we should expect and requires tuning.

# Analysis Report

## SSQ Table

Percent Variance Captured by Regression Model

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

Comp This Total This Total

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

1 99.38 99.38 28.65 28.65

2 0.51 99.88 6.41 35.05

3 0.06 99.94 4.13 39.18

4 0.04 99.99 23.34 62.52

5 0.01 100.00 0.01 62.53

6 0.00 100.00 0.46 62.99

7 0.00 100.00 0.95 63.94

8 0.00 100.00 3.44 67.38

9 0.00 100.00 3.54 70.92

10 0.00 100.00 22.90 93.82

11 0.00 100.00 2.24 96.06

12 0.00 100.00 1.53 97.59

13 0.00 100.00 0.35 97.94

## Model Measurements

Linear regression model using

Principal Components Regression

Developed 07-Nov-2019 19:30:12.336

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 185 (kebro@DESKTOP-B4UA9E4@20191107T190323.01335005 m:20191107190323.014)

Included: [ 1-15 ] [ 65-249 ]

Preprocessing: Mean Center

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191107T190328.99315403 m:20191107190328.995)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. PCs: 13

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

RMSEC: 0.0816029

RMSECV: 0.15795

RMSEP: 0.278471

Bias: -2.75335e-14

CV Bias: -0.00761313

Pred Bias:0.0561408

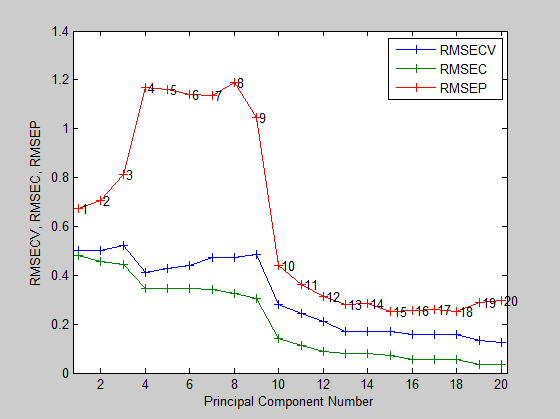
R^2 Cal: 0.979372

R^2 CV: 0.929845

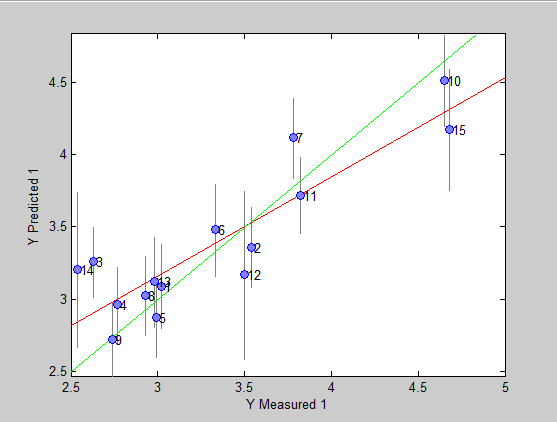
R^2 Pred: 0.836774

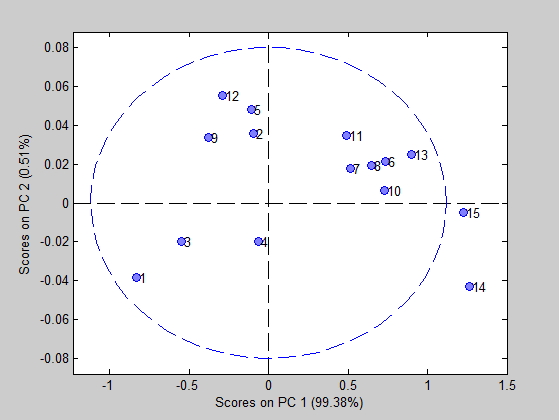
## Figures associated with the analysis:

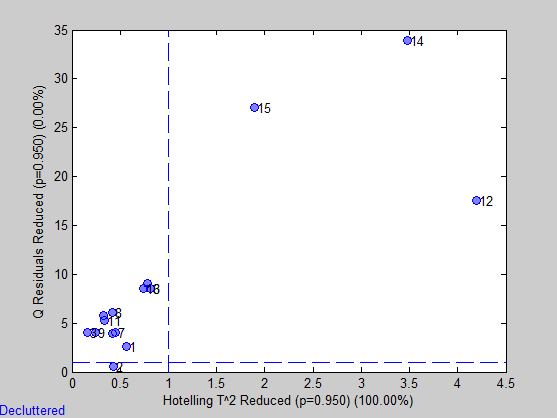
## Model Statistics - PCR 12 PCs - ,

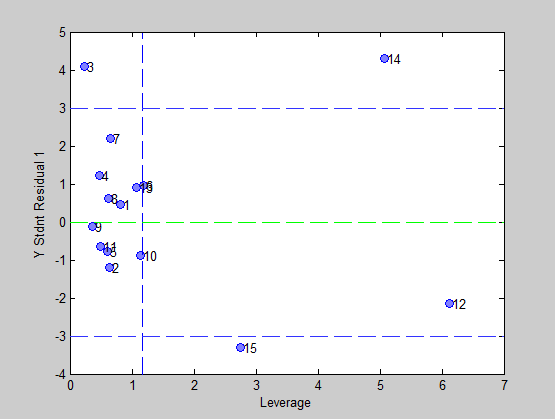


## Samples/Scores - PCR 12 PCs - ,

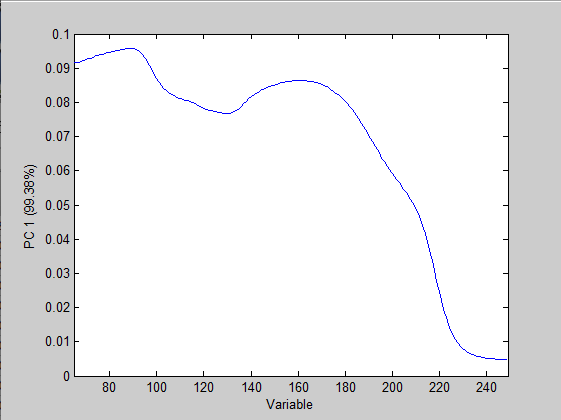




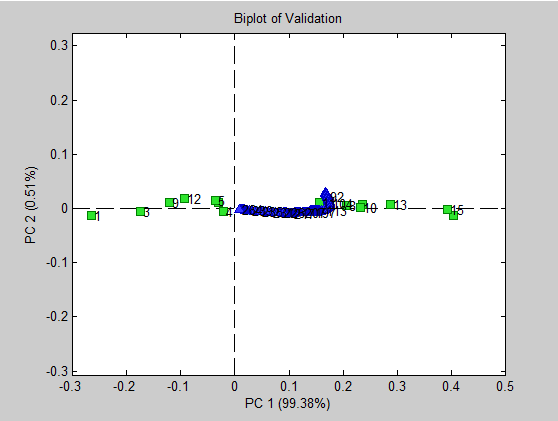




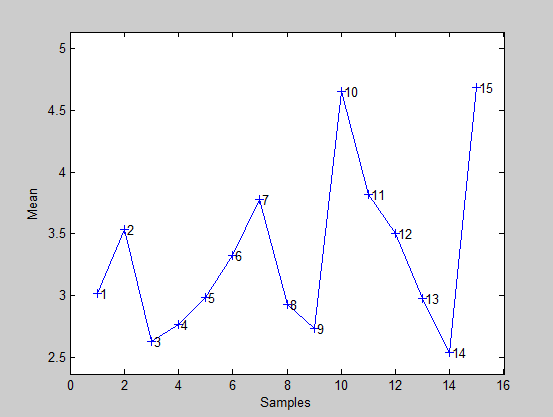
## Variables/Loadings - PCR 12 PCs - ,



## Biplot - PCR 12 PCs - ,



From the plots above, much like the PCA, we observe two distinct groupings appearing in the biplot with test sample 3, 12, 14, and 15 are showing some characteristics of being outliers. With this in mind we can say that our model generalizes to typical prediction data rather poorly as inate noise within each sample proves to be difficult to account for. The prediction biplot shows two distinct groups. Somewhat generally speaking, the groups are seperated by their protein content with lower values being on the left and higher values being on the right. The relation is not perfect as samples 13 and 14 appear grouped with the “high protein” group despite being on the lower end of the measured results.



# Application to New Data

Three samples were subsetted away from the training and validation sets and used as input to generate prediction results. We can say that samples one, two, and three have a protein content of approximately 4.6, 3, and 3.8 percent respectively. 95% prediction interval calculated use RMSEP.

Sample Y Predicted 1

1 4.628428091749 +- 0.5458032

2 3.029422544065 +- 0.5458032

3 3.738327011343 +- 0.5458032

# PLSR

Moving on to our partial least squares regression model we achieve comparable results to the PCR model with only 9 latent variables. Including more 9 LV’s does not prove to provide a significantly lower measurement of RMSECV. Similar pre-processing, data selection, and cross-validation methods were used in both models for the reasons as stated above. Within the summary and plots below, we see our model’s performance metrics are quite strong with improvements to each of the RMSE, R^2, and bias calculations over the PCR model. The generated biplot is once again showing two distinct groupings located in opposite areas of the plot. Cross-referencing against the measured Y-variables it appears the groupings represent high and low protein content. The SIMPLS

## SSQ Table

Percent Variance Captured by Regression Model

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

Comp This Total This Total

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

1 99.37 99.37 28.74 28.74

2 0.49 99.87 11.88 40.62

3 0.06 99.93 21.85 62.47

4 0.05 99.99 0.41 62.88

5 0.00 99.99 3.56 66.44

6 0.01 100.00 3.83 70.27

7 0.00 100.00 9.85 80.12

8 0.00 100.00 12.95 93.07

9 0.00 100.00 4.82 97.89

## Model Measurements

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 07-Nov-2019 19:37:38.023

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 185 (kebro@DESKTOP-B4UA9E4@20191107T190323.01335005 m:20191107190323.014)

Included: [ 1-15 ] [ 65-249 ]

Preprocessing: Mean Center

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191107T190328.99315403 m:20191107190328.995)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 9

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

RMSEC: 0.0824772

RMSECV: 0.187648

RMSEP: 0.266609

Bias: -3.10862e-14

CV Bias: -0.0191134

Pred Bias:0.0507799

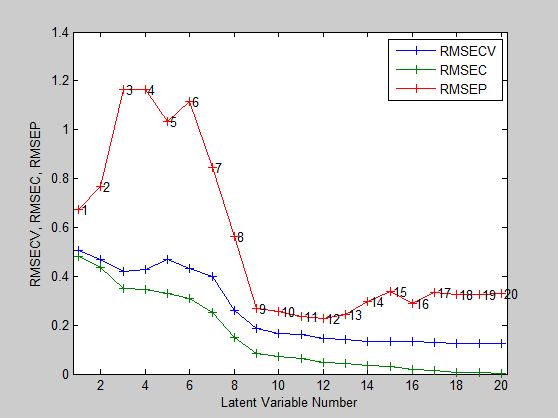
R^2 Cal: 0.978928

R^2 CV: 0.900946

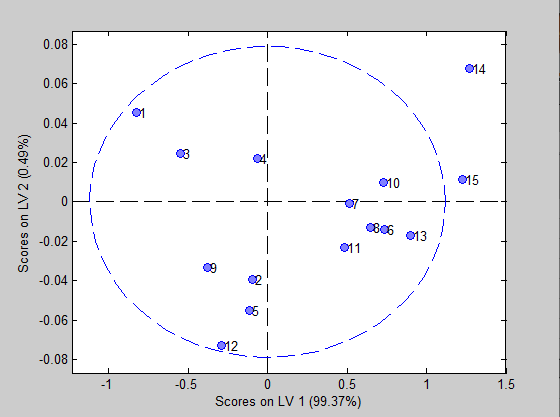
R^2 Pred: 0.851863

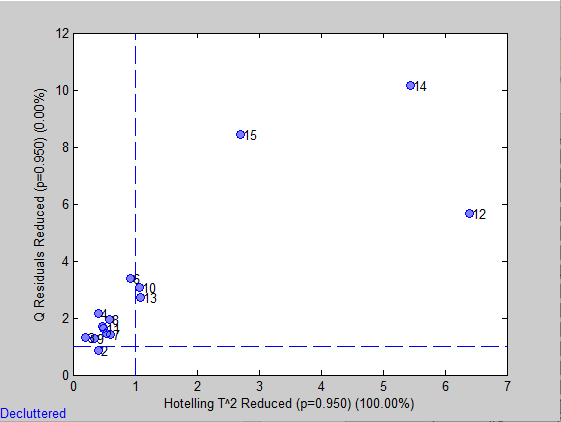
## Figures associated with the analysis:

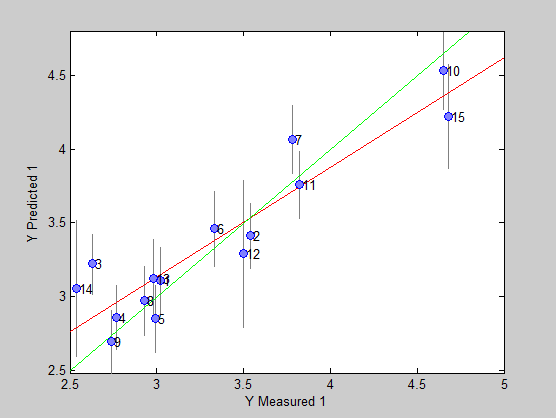
## Model Statistics - PLS 9 LVs - ,

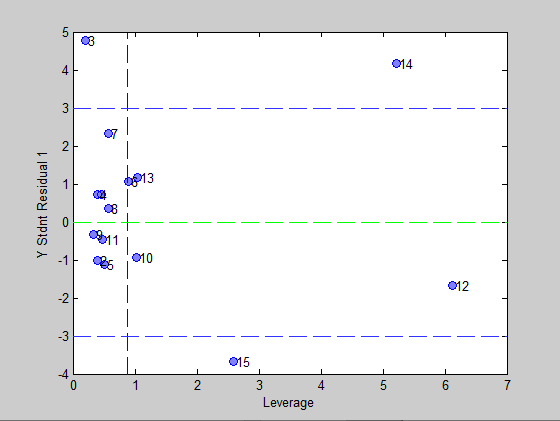


## Samples/Scores - PLS 9 LVs - ,

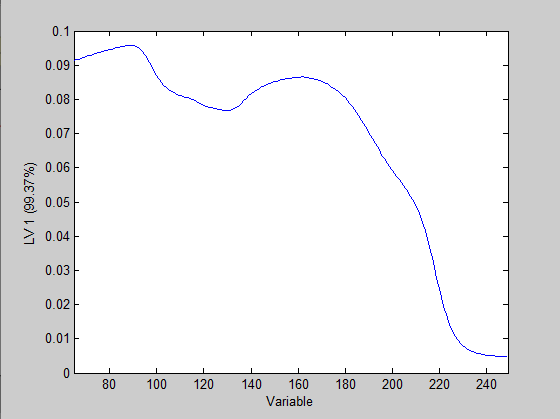




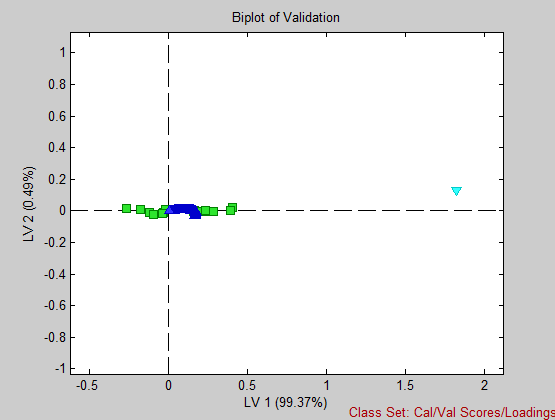




## Variables/Loadings - PLS 9 LVs - ,



## Biplot - PLS 9 LVs - ,



Here from the stdnt residual/leverage and Y-predicted/measured plots we see samples 3, 12, 14, and 15 from the validation set have the showings outliers. Much like with the PCR model, these observations are likely negatively impacting our prediction results quite significantly. Despite these samples, the PLSR model does preform well. Achieving a RMSECV of 0.193 and an R^2 CV of 0.901. The prediction results are also quite strong with an R^2 of 0.852 and a RMSE of 0.267.

# Application to New Data

3 samples were subsetted away from the training and validation sets and used as input to generate prediction results. Much like the PCR model, we can say that samples one, two, and three have a protein content of approximately 4.6, 3.0, and 3.7 percent respectively. 95% predicition intervals calculated using the RMSEP.

Sample Y Predicted 1

1 4.635776308620 +- 0.5225536

2 3.009235849203 +- 0.5225536

3 3.742637844354 +- 0.5225536

# Conclusions

We see that PCR and PLSR both perform adequately for the purposes of reducing high dimensional data and providing predictive models. The PCR and PLSR model produced