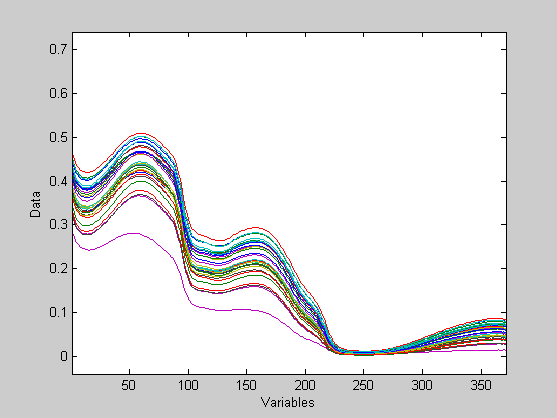
# Chemometrics Assignment 1 Kendall Brown, r0773111 KU Leuven, Session 1, 2019

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.

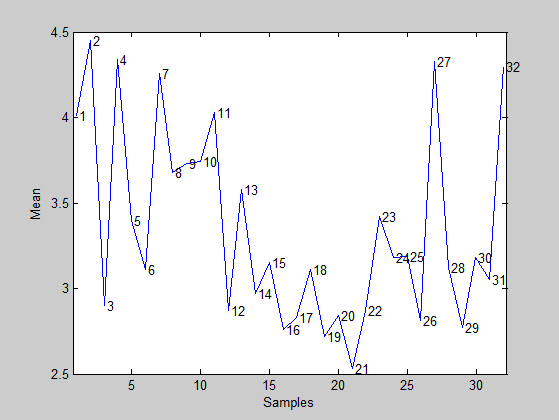
Regarding the dimensionality of the data. It is clear that given our data set it would be impractical to create a regression model without first conducting a dimensionality reduction. We have over 350 spectra measurements for each of our 32 samples in the calibration set. This is an issue as without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

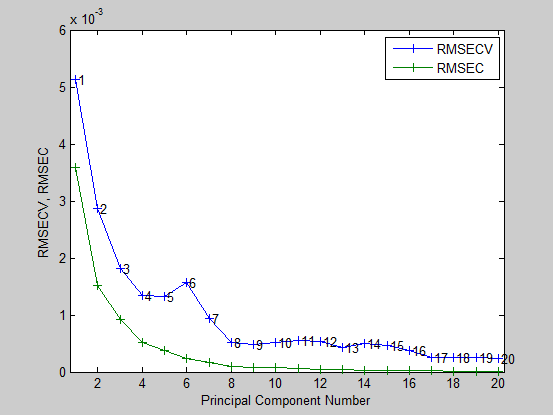
Preprocessing: Mean Center

Num. PCs: 8

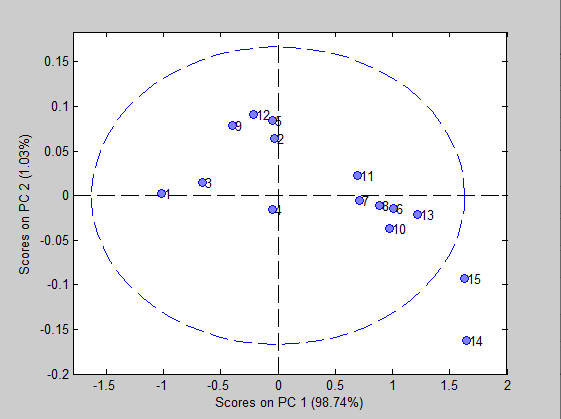
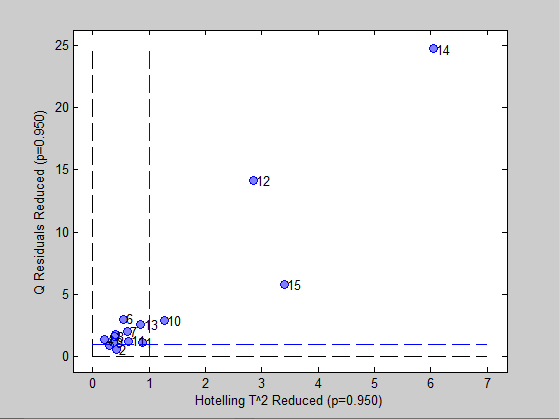
Algorithm: SVD

## Figures associated with the analysis:

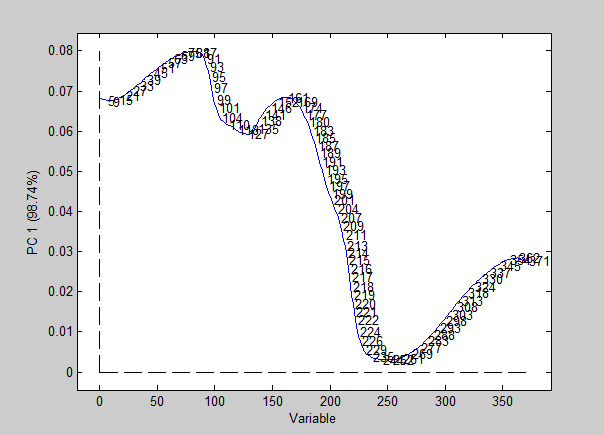
## Model Statistics - PCA 8 PCs -



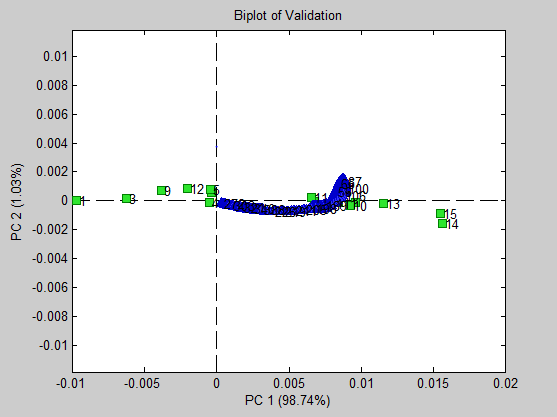
## Samples/Scores - PCA 8 PCs -



## Variables/Loadings - PCA 8 PCs -



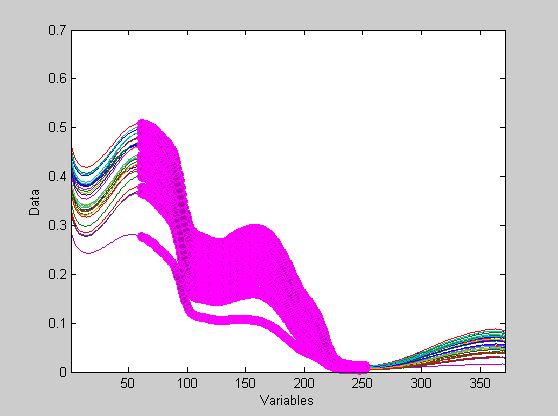
## Biplot - PCA 8 PCs -



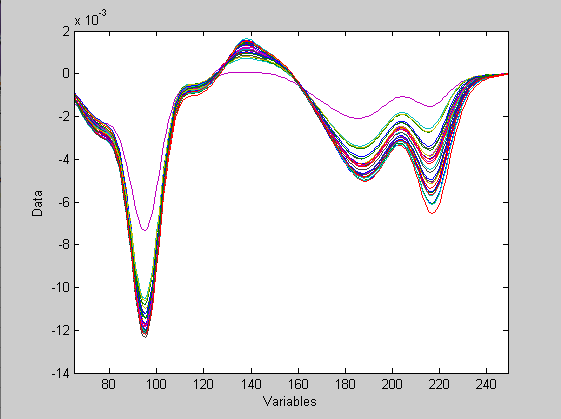
As observed in the above plots and summary statistics, we see that nearly all the information can be described in within the first component (explains 98.74% of observed variance). Additionally, we see that an 8-component model describes the data quite well as the RMSECV tapers off after the addition 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 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. To further emphasize the factors’ influence we employ a first-derivative pre-processing method. As seen below, the data loses quite a bit of unwanted noise and we observe a much more apparent change in spectra.



After applying a first-derivative pre-processing method, we observe that much of the noise has been removed and there is a much clearer representation of change in spectra. Additionally, we see noise being added into our data set by a singular outlier in sample 5. Sample 5 will be removed from the data set for the purpose of this analysis.

The results of the analysis show that an 11-component regression model carries strong validation and prediction results. We calculated a RMSE of cross-validation to approximately 0.161, a R^2 of cross-validation of 0.931, and a cross-validation bias of -0.019. Similarly, we see our model produces prediction results comparable to our validation. We see a RMSE of prediction of 0.229, a R^2 prediction of 0.898 and a prediction bias of 0.097.

# Analysis Report

## Model

Linear regression model using

Principal Components Regression

Developed 06-Nov-2019 16:55:44.347

Author: kebro@DESKTOP-B4UA9E4

X-block: 31 by 185 (kebro@DESKTOP-B4UA9E4@20191105T130418.07789196 m:20191106165507.179)

Included: [ 1-4 6-32 ] [ 65-249 ]

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

Y-block: 31 by 1 (kebro@DESKTOP-B4UA9E4@20191106T165426.43737714 m:20191106165507.165)

Included: [ 1-4 6-32 ] [ 1 ]

Preprocessing: Autoscale

Num. PCs: 11

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

RMSEC: 0.0867252

RMSECV: 0.161176

Bias: 4.70747e-05

CV Bias: -0.0186288

R^2 Cal: 0.976701

R^2 CV: 0.930651

Percent Variance Captured by Regression Model

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

Comp This Total This Total

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

1 99.40 99.40 0.11 0.11

2 0.58 99.97 35.73 35.84

3 0.02 99.99 2.42 38.26

4 0.01 99.99 6.25 44.51

5 0.00 100.00 16.06 60.57

6 0.00 100.00 0.04 60.62

7 0.00 100.00 0.16 60.78

8 0.00 100.00 20.55 81.33

9 0.00 100.00 13.46 94.78

10 0.00 100.00 0.95 95.73

11 0.00 100.00 1.94 97.67

## Prediction

Linear regression model using

Principal Components Regression

Developed 06-Nov-2019 16:56:28.287

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 185 (kebro@DESKTOP-B4UA9E4@20191106T165616.99503253 m:20191106165616.997)

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

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

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191106T165626.07064010 m:20191106165626.073)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. PCs: 11

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

RMSEC: 0.0867252

RMSECV: 0.161176

RMSEP: 0.229443

Bias: 4.70747e-05

CV Bias: -0.0186288

Pred Bias:0.0974724

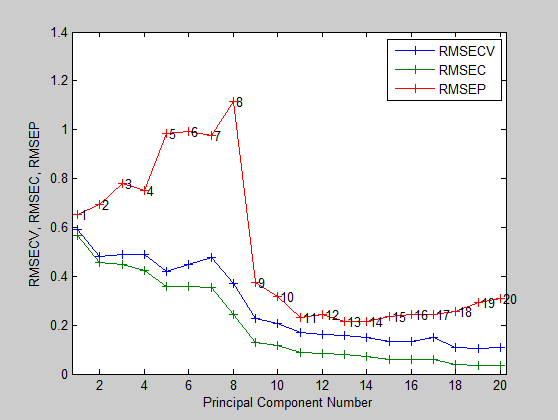
R^2 Cal: 0.976701

R^2 CV: 0.930651

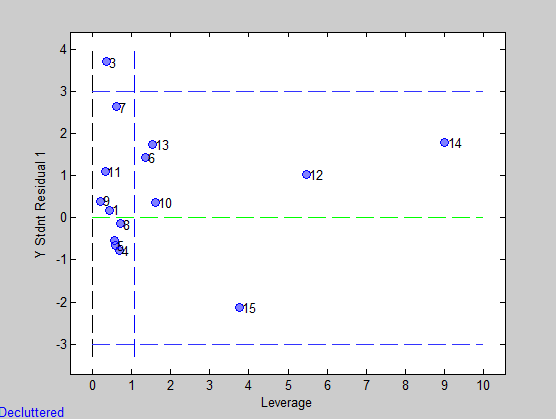
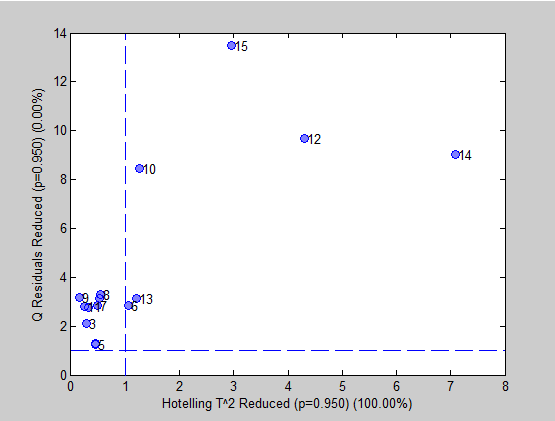
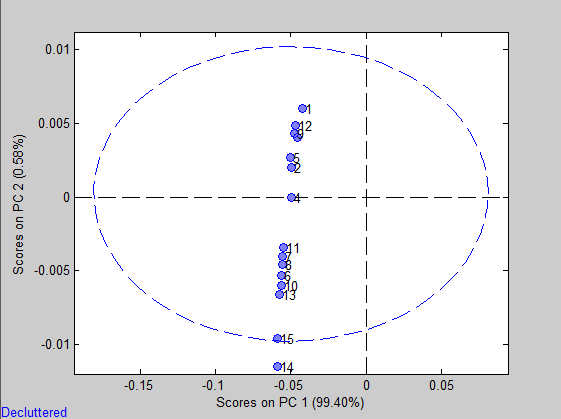
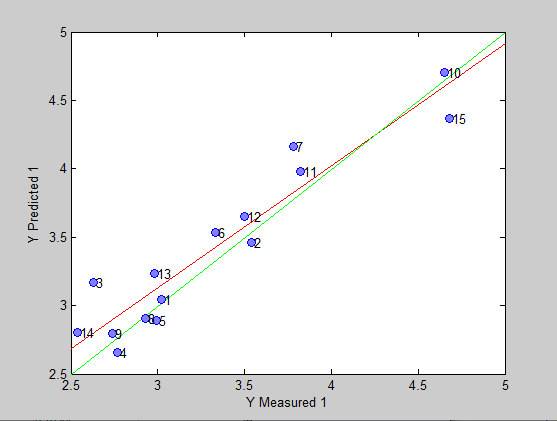
R^2 Pred: 0.898166

## Figures associated with the analysis:

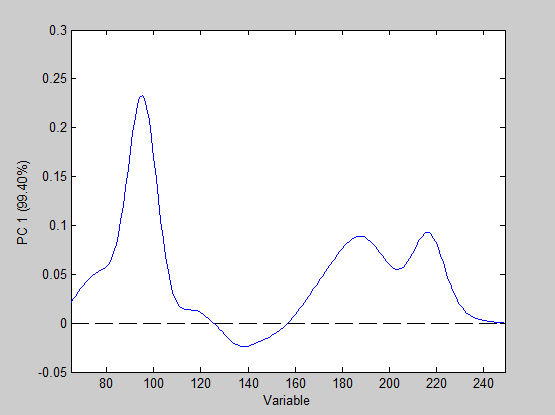
## Model Statistics - PCR 11 PCs - ,



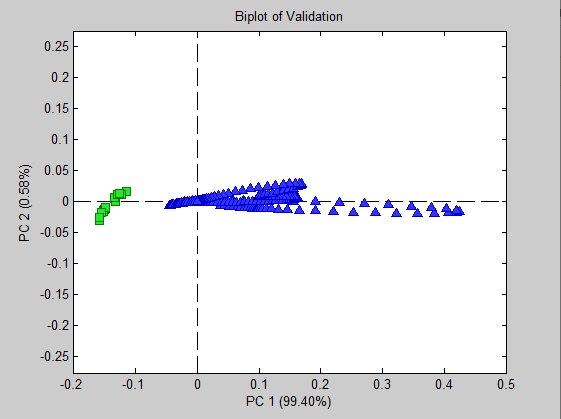
## Samples/Scores - PCR 11 PCs - ,



## Variables/Loadings - PCR 11 PCs - ,



## Biplot - PCR 11 PCs - ,



From the plots above, much like the PCA, we observe two distinct groupings appearing in the biplot with test sample 3 showing some characteristics of being an outlier. Not surprisingly this sample is does influence our prediction results negatively as shown in the table below. With this in mind we can say that our model generalizes to typical prediction data rather well with RMSEs and R^2s of CV and prediction to be practically identical.

Linear regression model using

Principal Components Regression

Developed 05-Nov-2019 16:00:32.696

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 185 (kebro@DESKTOP-B4UA9E4@20191105T132642.80753425 m:20191105132642.807)

Included: [ 1-2 4-15 ] [ 65-249 ]

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

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191105T13302.54260554 m:20191105153439.400)

Included: [ 1-2 4-15 ] [ 1 ]

Preprocessing: Autoscale

Num. PCs: 11

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

RMSEC: 0.0867252, RMSECV: 0.161169

RMSEP: 0.188923, Bias: 4.70747e-05

CV Bias: -0.01335

Pred Bias:4.70747e-05

R^2 Cal: 0.976701

R^2 CV: 0.930927

R^2 Pred: 0.924896

# Application to New Data

3 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, 2.9, and 3.5 percent respectively.

Sample Y Predicted 1

1 4.598139207264

2 2.916376276049

3 3.549356649662

Moving on to our partial least squares regression model we achieve comparable results to the PCR model with only 9 latent variables. Similar pre-processing, data selection, and cross-validation methods were used in both models for the same 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.

## Model

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 05-Nov-2019 13:42:50.799

Author: kebro@DESKTOP-B4UA9E4

X-block: datatrimmed 31 by 185 (kebro@DESKTOP-B4UA9E4@20191105T130418.07789196 m:20191105133110.797)

Included: [ 1-4 6-32 ] [ 65-249 ]

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

Y-block: 31 by 1 (kebro@DESKTOP-B4UA9E4@20191105T130430.01163904 m:20191105130806.254)

Included: [ 1-4 6-32 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 9

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

RMSEC: 0.0746788

RMSECV: 0.149935

Bias: -4.51241e-06

CV Bias: -0.0141281

R^2 Cal: 0.982724

R^2 CV: 0.937907

## SSQ Table

Percent Variance Captured by Regression Model

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

Comp This Total This Total

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

1 98.31 98.31 0.92 0.92

2 1.66 99.97 35.43 36.35

3 0.01 99.98 19.53 55.88

4 0.01 99.99 5.90 61.78

5 0.01 100.00 2.20 63.98

6 0.00 100.00 27.04 91.02

7 0.00 100.00 3.36 94.39

8 0.00 100.00 1.39 95.78

9 0.00 100.00 2.50 98.27

## Prediction

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 05-Nov-2019 13:42:51.033

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 185 (kebro@DESKTOP-B4UA9E4@20191105T132642.80753425 m:20191105132642.807)

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

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

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191105T132651.54115199 m:20191105132651.556)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 9

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

RMSEC: 0.0746788

RMSECV: 0.149935

RMSEP: 0.20351

Bias: -4.51241e-06

CV Bias: -0.0141281

Pred Bias:0.0454985

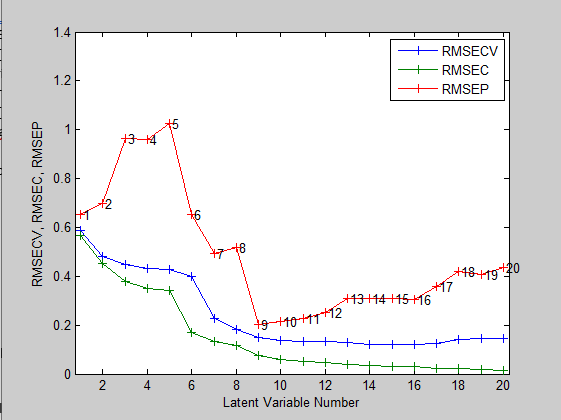
R^2 Cal: 0.982724

R^2 CV: 0.937907

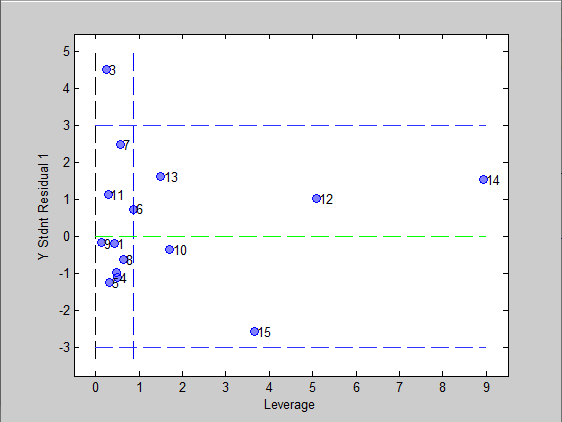
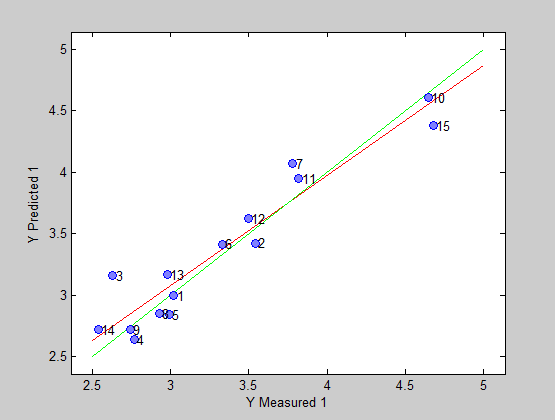
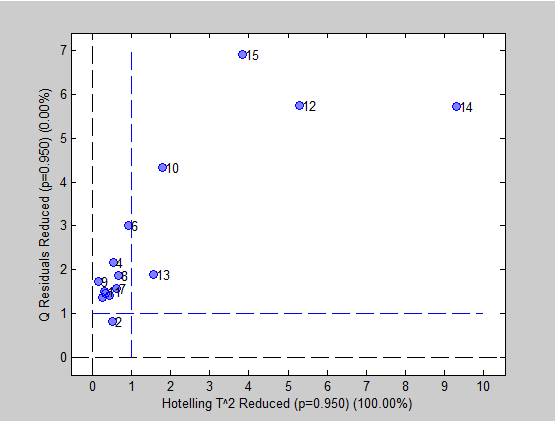
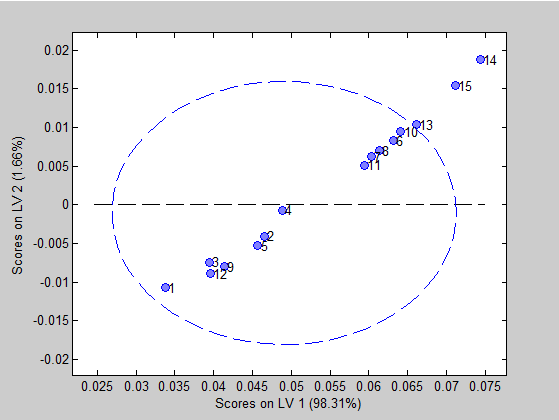
R^2 Pred: 0.907251

## Figures associated with the analysis:

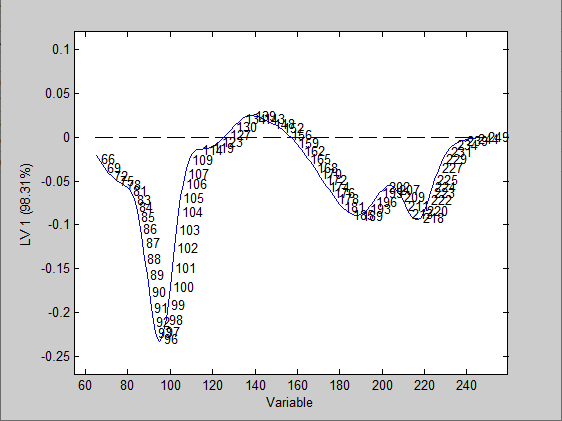
## Model Statistics - PLS 9 LVs - ,



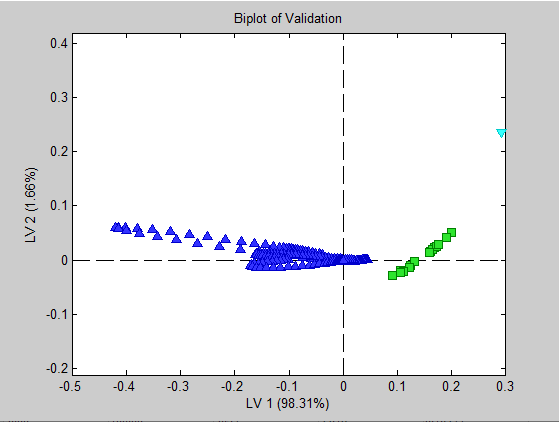
## Samples/Scores - PLS 9 LVs - ,



## Variables/Loadings - PLS 9 LVs - ,



## Biplot - PLS 9 LVs - ,



Here from the stdnt residual/leverage plot we see sample 3 from the validation set has the showings of an outlier. Much like with the PCR model, this observation is negatively impacting our prediction results quite significantly. Removing it shows that our model generalizes to typical data quite well.

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 06-Nov-2019 18:09:46.688

Author: kebro@DESKTOP-B4UA9E4

X-block: 14 by 185 (kebro@DESKTOP-B4UA9E4@20191106T165616.99503253 m:20191106170301.873)

Included: [ 1-2 4-15 ] [ 65-249 ]

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

Y-block: 14 by 1 (kebro@DESKTOP-B4UA9E4@20191106T165626.07064010 m:20191106170301.850)

Included: [ 1-2 4-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 9

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

RMSEC: 0.0746788

RMSECV: 0.15375

RMSEP: 0.156337

Bias: -4.51241e-06

CV Bias: -0.0153318

Pred Bias:0.011015

R^2 Cal: 0.982724

R^2 CV: 0.932424

R^2 Pred: 0.941627

# 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, 2.9, and 3.5 percent respectively.

Sample Y Predicted 1

1 4.616021266063

2 2.874367326164

3 3.512197190220

# Conclusions

We see that PCR and PLSR both perform well for the purposes of reducing high dimensional data and providing predictive models. PLSR does outperform PCR by being able to produce an equivalently powerful model using two fewer inputs. Considering the initial size of the data set, it is fair to say both methods provide adequate results.