### Kendall Brown r0773111 Chemometrics, Fall 2019 Assignment 2

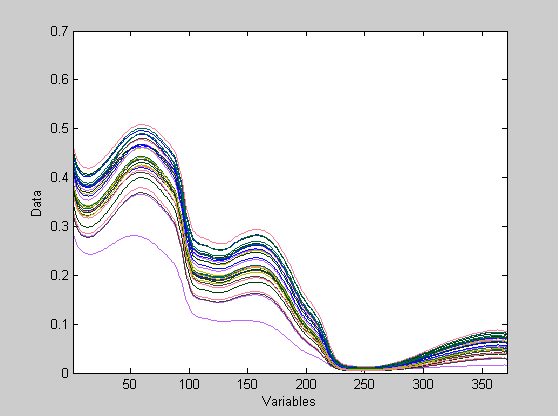
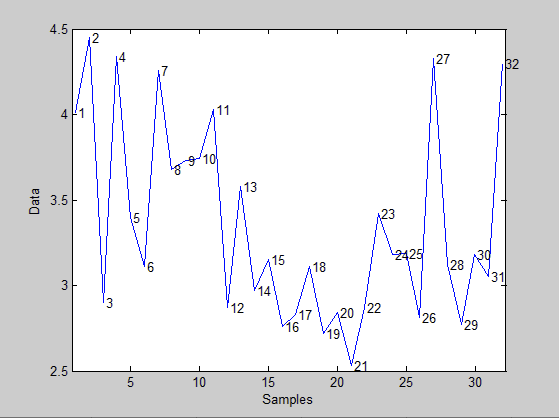
# Task

Compare and contrast different pre-processing methods which are to be applied to a PLS regression model. The PLSR model will be built using milk reflectance spectra vs protein content data. No variable selection will be performed as that is the focus of assignment 3.

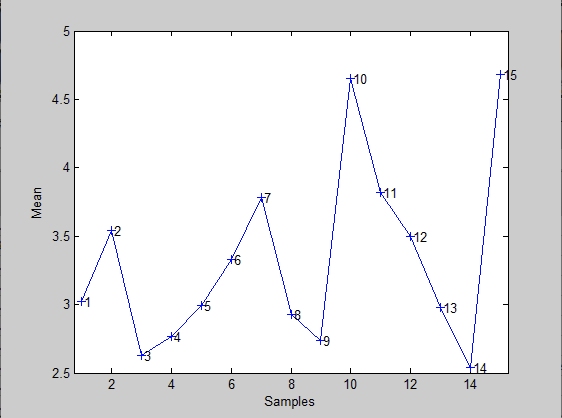
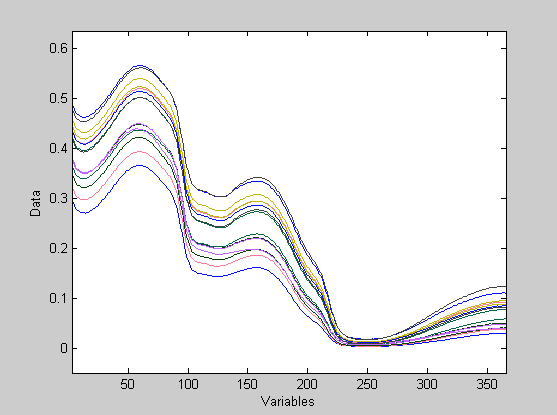
# Data

Before we begin a comparison, we must determine which methods would provide the best results. Shown below, the spectra of both the calibration and test data can be seen to have general trend with some noise. Through the application of pre-processing methods, our goal is to exaggerate the relationship between the measured spectra and the measured protein content of a given milk sample.

Calibration:

Test:



# Outliers

Based on results from this report, observation 5 of the calibration data and observations 12 and 14 from the test data appear to have a profoundly detrimental effect on the building and evaluation of our model. Because take home 3 focuses on variable selection and model trimming, their presence will be addressed there.

# Baseline

To begin the comparison, it is important to establish a baseline model for which all other methods will be measured against. For this, we shall use a model for which no pre-processing method has been employed. Found in the tables below will be the results of a 12 latent variable model. As can be seen in the measured statistics the model performs decently. With an RMSECV of 0.227 and an R^2 CV of 0.858, we can say that the model generally generalizes to itself quite well. Unfortunately, we can not say the same for new data. With an RMSEP of 0.576 and an R^2 Pred of 0.434 we can say quite confidently that the model does not reliably provide strong predication results. This will be addressed later as more pre-processing methods are applied to the model.

## SSQ Table

Percent Variance Captured by Regression Model

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

Comp This Total This Total

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

1 99.83 99.83 0.42 0.42

2 0.16 99.99 30.80 31.22

3 0.01 100.00 7.40 38.61

4 0.00 100.00 17.89 56.51

5 0.00 100.00 12.02 68.53

6 0.00 100.00 4.79 73.32

7 0.00 100.00 4.01 77.32

8 0.00 100.00 3.94 81.27

9 0.00 100.00 11.23 92.50

10 0.00 100.00 3.18 95.67

11 0.00 100.00 2.25 97.92

12 0.00 100.00 0.24 98.15

## Prediction

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 13-Nov-2019 20:55:43.140

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 371 (kebro@DESKTOP-B4UA9E4@20191113T193453.60567726 m:20191113193453.608)

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

Preprocessing: None

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191113T193500.97414935 m:20191113193500.977)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 12

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

RMSEC: 0.0759916

RMSECV: 0.226776

RMSEP: 0.575861

Bias: -8.05027e-06

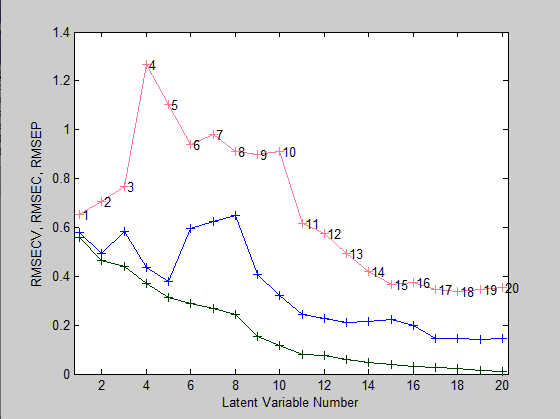
CV Bias: 0.0370657

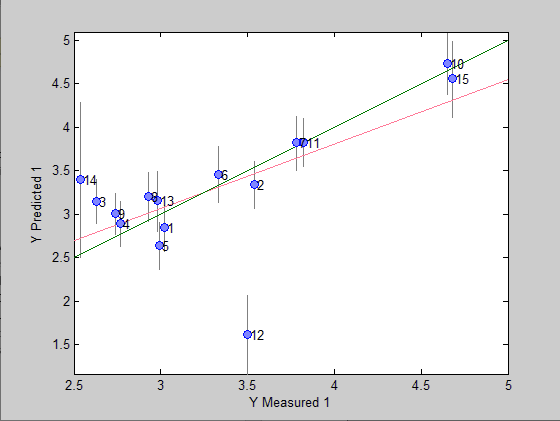
Pred Bias:-0.0155544

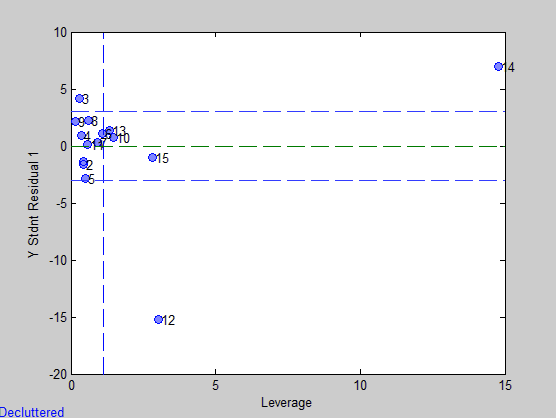
R^2 Cal: 0.981539

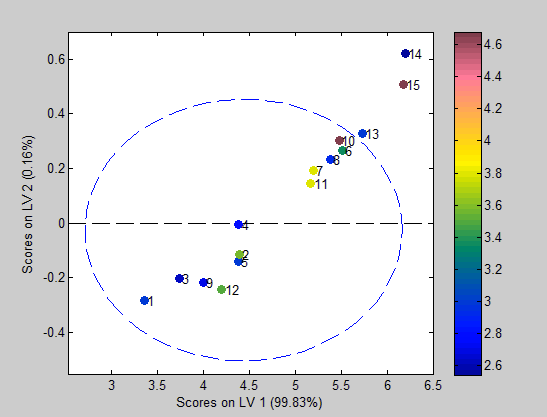
R^2 CV: 0.858346

R^2 Pred: 0.43395



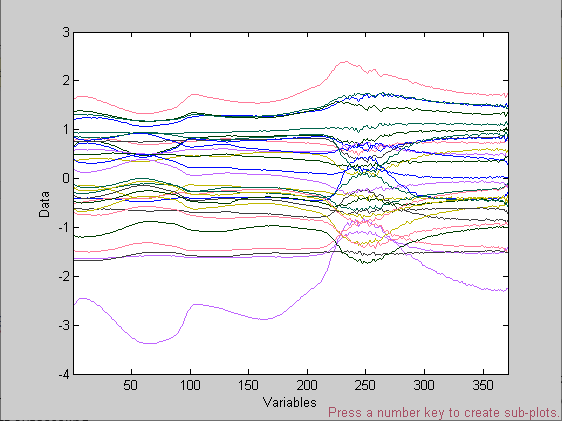




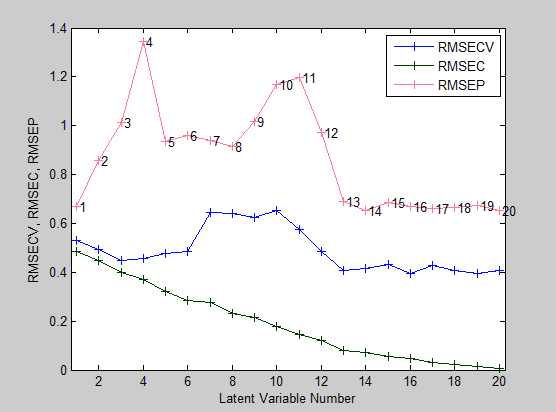


# Scaling

By default, Solo incorporates their Autoscaling pre-processing function. This is a normalization method which seeks to level out the spectra within its given domain. Seen Below is a visualization of the spectra post autoscaling.



Unfortunately, we are seeing the autoscaling function injecting a lot of noise into our spectra. The results of the analysis show this quite well as our model suffers rather significantly. We see a three latent variable model could be the best of the models we could use. However, the RMSE and R^2 values of both the cross-validation and prediction are quite a bit worse than the model with no pre-processing. The R^2 of cross validation is below any reasonable expectation of adequacy, and the R^2 of prediction might as well be zero. We will have to use more advanced methods of pre-processing before building our final model.



## SSQ Table

Percent Variance Captured by Regression Model

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

Comp This Total This Total

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

1 95.90 95.90 24.19 24.19

2 3.08 98.98 11.46 35.65

3 0.84 99.82 13.46 49.10

## Prediction

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 14-Nov-2019 15:04:12.356

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 371 (kebro@DESKTOP-B4UA9E4@20191114T150401.70766003 m:20191114150401.709)

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

Preprocessing: Autoscale

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191114T150407.33437606 m:20191114150407.336)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 3

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

RMSEC: 0.399007

RMSECV: 0.448085

RMSEP: 1.01398

Bias: 4.44089e-16

CV Bias: -0.0043399

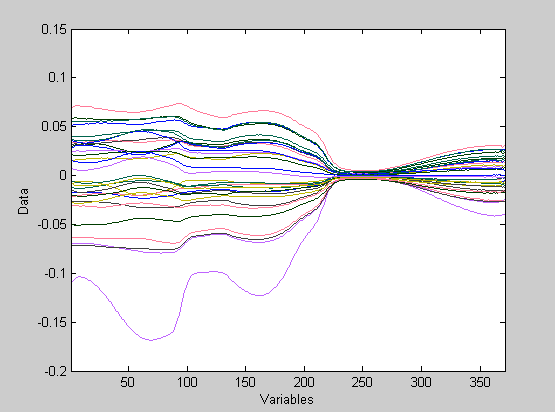
Pred Bias:0.356197

R^2 Cal: 0.491042

R^2 CV: 0.373361

R^2 Pred: 0.0347904

An alternative scaling method would be a mean-centering operation. Doing so results in decent model power measurements as can be seen below.



## SSQ Table

Percent Variance Captured by Regression Model

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

Comp This Total This Total

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

1 98.73 98.73 19.82 19.82

2 1.04 99.77 18.83 38.65

3 0.11 99.88 13.52 52.17

4 0.09 99.97 5.62 57.79

5 0.01 99.98 12.23 70.02

6 0.01 99.99 4.72 74.74

7 0.01 100.00 2.31 77.05

8 0.00 100.00 13.16 90.21

9 0.00 100.00 4.43 94.64

10 0.00 100.00 1.79 96.43

11 0.00 100.00 1.30 97.73

## Prediction

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 14-Nov-2019 21:45:43.930

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 371 (kebro@DESKTOP-B4UA9E4@20191114T153717.93610099 m:20191114154515.881)

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

Preprocessing: Mean Center

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191114T153725.97610212 m:20191114154515.867)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 11

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

RMSEC: 0.0842064

RMSECV: 0.206176

RMSEP: 0.603712

Bias: -1.06581e-14

CV Bias: -0.00447995

Pred Bias:-0.0772996

R^2 Cal: 0.977332

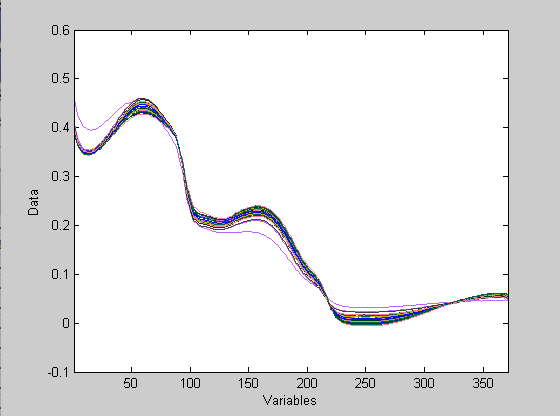
R^2 CV: 0.886025

R^2 Pred: 0.380454

Applying a mean-centering pre-processing method results in a cross-validation R^2 of .886 and an RMSE of .206. The prediction results leave something to be desired as they are being negatively impacted by outliers. Overall it appears mean-centering allows for an 11 component model with a good amount of power.

# Weighting

To denoise our spectra, we will be using a multiplicative scatter correction (MSC) taking the mean as the support. We see this to be rather effective as much of the noise is seemingly eliminated. The change in spectra is as noticeable as it was in the unaltered spectra. We do achieve some rather strong results for our cross validation with an R^2 of .92 and an RMSE of .23. Our predication metrics are unfortunately being influenced by outliers. This will be addressed during model trimming and variable selection.



## SSQ Table

Percent Variance Captured by Regression Model

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

Comp This Total This Total

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

1 1.07 1.07 23.05 23.05

2 98.92 99.99 0.23 23.28

3 0.01 100.00 9.32 32.60

4 0.00 100.00 27.81 60.40

5 0.00 100.00 8.42 68.83

6 0.00 100.00 3.45 72.28

7 0.00 100.00 4.41 76.69

8 0.00 100.00 14.40 91.09

9 0.00 100.00 3.71 94.80

10 0.00 100.00 2.64 97.44

11 0.00 100.00 0.50 97.94

12 0.00 100.00 0.38 98.32

## Prediction

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 14-Nov-2019 16:03:55.258

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 371 (kebro@DESKTOP-B4UA9E4@20191114T153717.93610099 m:20191114154515.881)

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

Preprocessing: MSC (mean)

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191114T153725.97610212 m:20191114154515.867)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 12

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

RMSEC: 0.0724849

RMSECV: 0.227643

RMSEP: 0.594879

Bias: -4.01476e-09

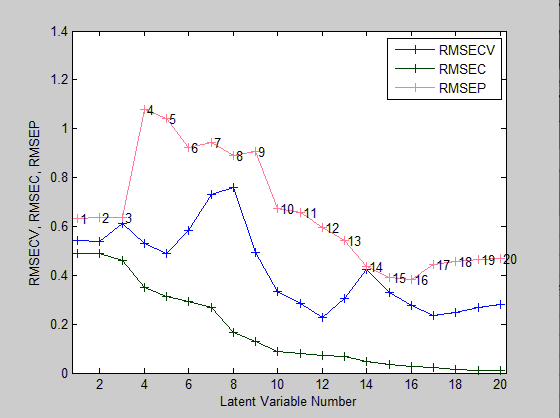
CV Bias: -0.00151499

Pred Bias:0.12026

R^2 Cal: 0.983204

R^2 CV: 0.916517

R^2 Pred: 0.392515



An alternative to a MSC is a standard normal variate weighting (SNV). Here we are converting the data into normalized values. As seen below we achieve a structure very similar to the MSC pre-processed data. As expected, we achieve similar results as well.

## SSQ Table

Percent Variance Captured by Regression Model

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

Comp This Total This Total

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

1 5.09 5.09 22.30 22.30

2 94.88 99.97 1.15 23.44

3 0.03 100.00 9.22 32.66

4 0.00 100.00 27.76 60.43

5 0.00 100.00 8.41 68.83

6 0.00 100.00 3.45 72.28

7 0.00 100.00 4.41 76.70

8 0.00 100.00 14.39 91.08

9 0.00 100.00 3.73 94.81

10 0.00 100.00 2.63 97.44

11 0.00 100.00 0.50 97.94

12 0.00 100.00 0.38 98.32

## Prediction

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 14-Nov-2019 17:21:19.674

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 371 (kebro@DESKTOP-B4UA9E4@20191114T153717.93610099 m:20191114154515.881)

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

Preprocessing: SNV

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191114T153725.97610212 m:20191114154515.867)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 12

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

RMSEC: 0.0724463

RMSECV: 0.238197

RMSEP: 0.593961

Bias: -1.98578e-06

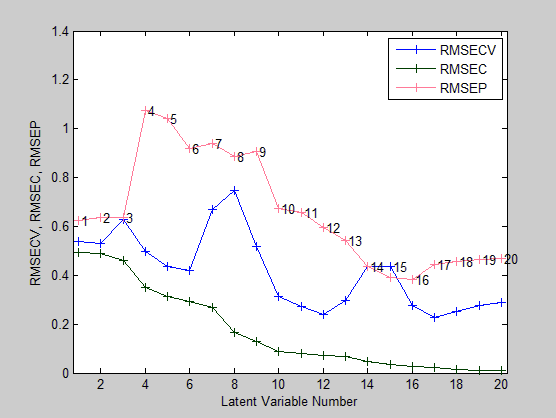
CV Bias: -0.0144263

Pred Bias:0.11954

R^2 Cal: 0.983221

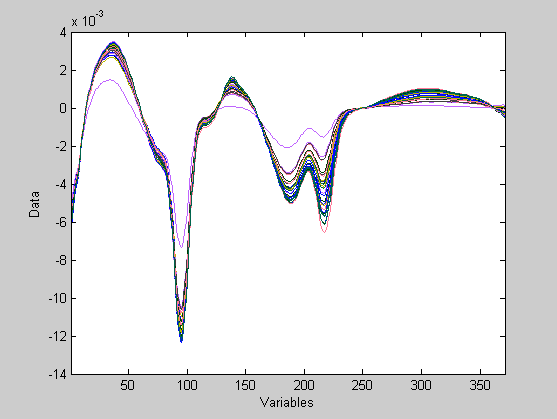
R^2 CV: 0.897748

R^2 Pred: 0.392925



# Filtering

For our filtering methods we will be taking derivatives across the spectra. We will start with a 1st order derivative. As seen below, applying a 1st order derivative exaggerates the change in spectra quite substantially whilst reducing much of the troublesome noise. The resulting model shows considerable improvement. An eleven latent variable model provides a cross validation R^2 of .929 and an RMSE of .182.



## SSQ Table

Percent Variance Captured by Regression Model

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

Comp This Total This Total

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

1 97.39 97.39 1.13 1.13

2 2.53 99.91 32.19 33.32

3 0.04 99.95 11.59 44.91

4 0.03 99.98 9.11 54.02

5 0.01 99.99 10.93 64.95

6 0.00 100.00 7.40 72.35

7 0.00 100.00 13.41 85.76

8 0.00 100.00 6.31 92.07

9 0.00 100.00 3.29 95.36

10 0.00 100.00 2.14 97.50

11 0.00 100.00 1.20 98.70

## Prediction

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 14-Nov-2019 18:29:26.984

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 371 (kebro@DESKTOP-B4UA9E4@20191114T153717.93610099 m:20191114154515.881)

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

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

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191114T153725.97610212 m:20191114154515.867)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 11

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

RMSEC: 0.0637473

RMSECV: 0.182228

RMSEP: 0.474025

Bias: -0.000196723

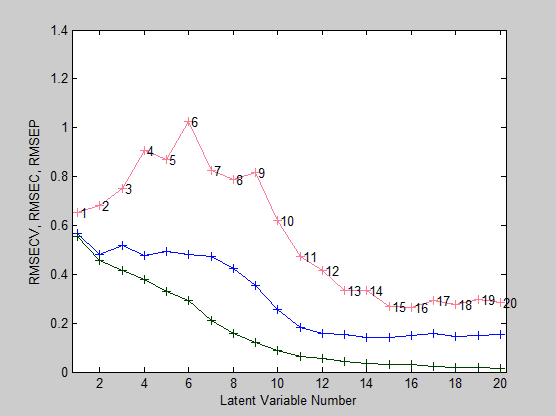
CV Bias: -0.00183455

Pred Bias:0.0413312

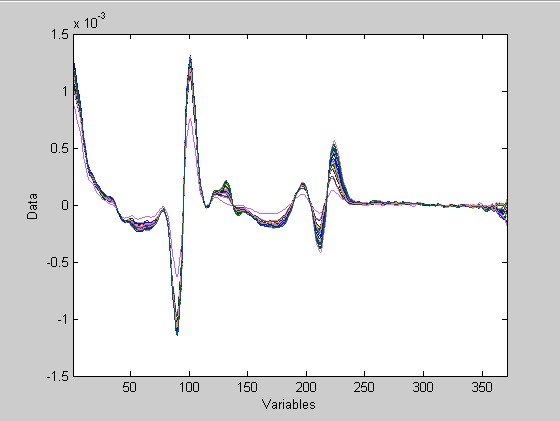
R^2 Cal: 0.987009

R^2 CV: 0.928583

R^2 Pred: 0.576621



Increasing to a second order derivative yields weaker results. Although much of the noise has been removed from the data set, and the importance of certain factors becomes much more apparent, a second order derivative results in a cross-validation R^2 value of .836 and an RMSE of .247. Similarly, the predication results are weaker as well. For this data set it could be said that a first order derivative is a strong pre-processing method.



## SSQ Table

Percent Variance Captured by Regression Model

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

Comp This Total This Total

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

1 89.70 89.70 3.70 3.70

2 10.19 99.88 30.74 34.44

3 0.02 99.91 18.19 52.63

4 0.03 99.94 8.83 61.45

5 0.03 99.97 4.77 66.23

6 0.01 99.98 3.72 69.95

7 0.00 99.99 6.90 76.85

8 0.00 99.99 12.19 89.04

9 0.00 99.99 4.19 93.23

10 0.00 100.00 2.42 95.65

11 0.00 100.00 1.96 97.61

## Prediction

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 14-Nov-2019 19:49:38.372

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 371 (kebro@DESKTOP-B4UA9E4@20191114T153717.93610099 m:20191114154515.881)

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

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

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191114T153725.97610212 m:20191114154515.867)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 11

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

RMSEC: 0.0863953

RMSECV: 0.246609

RMSEP: 0.545071

Bias: 0.000125865

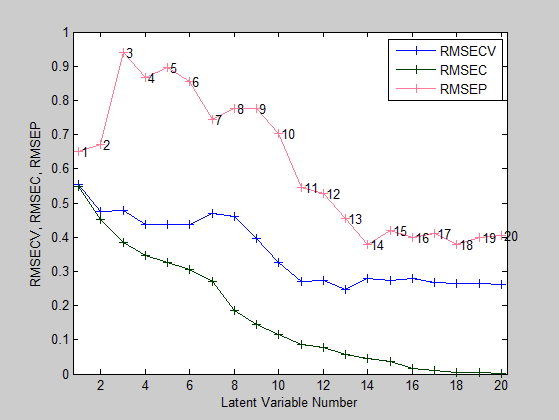
CV Bias: -0.0265718

Pred Bias:-0.200719

R^2 Cal: 0.976138

R^2 CV: 0.836423

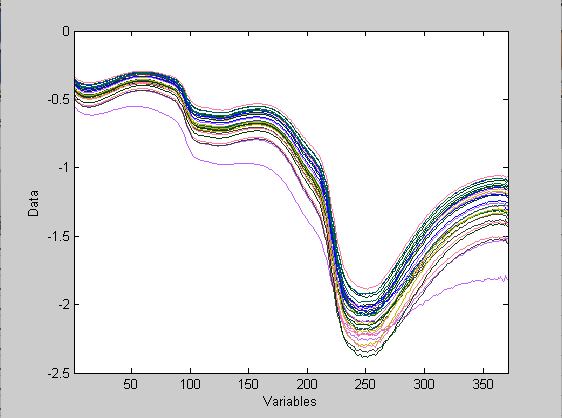
R^2 Pred: 0.472927



# Transformations

Transformations play a key role in data pre-processing. With the correct transforms it is possible to turn non-linear data linear. For this example, a log10 transform shall be applied to the data to determine if there exists any form of exponentiality.

Shown below are the results of the transform where we observe a surprising result. With a 1 latent variable model a CV R^2 of .596 and a RMSE of .575 is achieved. The predication and calibration results are unsurprisingly terrible. However, it should be noted that there do exist outliers in both sets which are most likely impacting the results in a negative way. Additionally, it appears the addition of more latent variables results in a less powerful model. This is an issue as we would expect to see a decreasing relationship between RMSE and the number of latent variables. Such a relation is likely the result of a log10 transform being a generally bad pre-processing method for this particular data set.



## SSQ Table

Percent Variance Captured by Regression Model

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

Comp This Total This Total

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

1 99.87 99.87 0.20 0.20

## Prediction

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 14-Nov-2019 21:17:26.727

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 371 (kebro@DESKTOP-B4UA9E4@20191114T153717.93610099 m:20191114154515.881)

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

Preprocessing: Log10

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191114T153725.97610212 m:20191114154515.867)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 1

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

RMSEC: 0.558743

RMSECV: 0.575102

RMSEP: 0.649434

Bias: -0.0246762

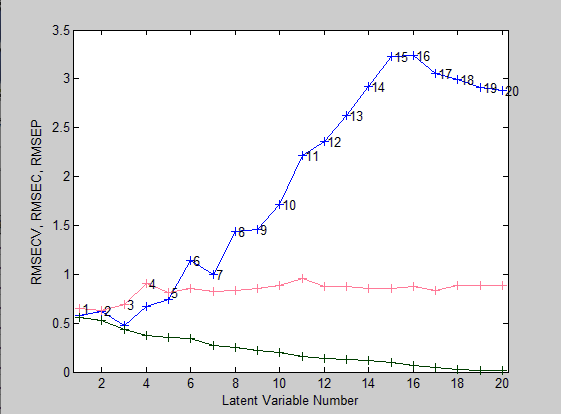
CV Bias: -0.0243731

Pred Bias:0.000936409

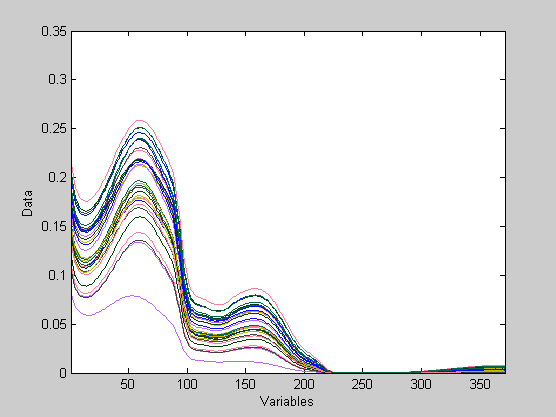
R^2 Cal: 0.211837

R^2 CV: 0.595761

R^2 Pred: 0.158866



Another transform which may provide additional model power would be a power transform. Here the data is taken to be raised to the power of a certain constant, in this case 2. Much of the noise at the far-right tale has been eliminated. The same cannot be said for the left-hand tail as it is quite clear there is a significant amount of noise present in those variables. From the results below we observe rather decent results. We calculate a cross-validation R^2 of .85 and RMSE of .226. Much like the other pre-processing methods prediction results are struggling with outliers.



## SSQ Table

Percent Variance Captured by Regression Model

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

Comp This Total This Total

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

1 99.83 99.83 1.03 1.03

2 0.16 99.98 32.92 33.95

3 0.01 99.99 15.21 49.15

4 0.01 100.00 13.74 62.89

5 0.00 100.00 8.31 71.21

6 0.00 100.00 4.23 75.44

7 0.00 100.00 5.89 81.32

8 0.00 100.00 11.03 92.35

9 0.00 100.00 3.48 95.83

## Prediction

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 14-Nov-2019 21:33:59.772

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 371 (kebro@DESKTOP-B4UA9E4@20191114T153717.93610099 m:20191114154515.881)

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

Preprocessing: power: x = x^2

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191114T153725.97610212 m:20191114154515.867)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 9

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

RMSEC: 0.114267

RMSECV: 0.225546

RMSEP: 1.09743

Bias: 0.000631017

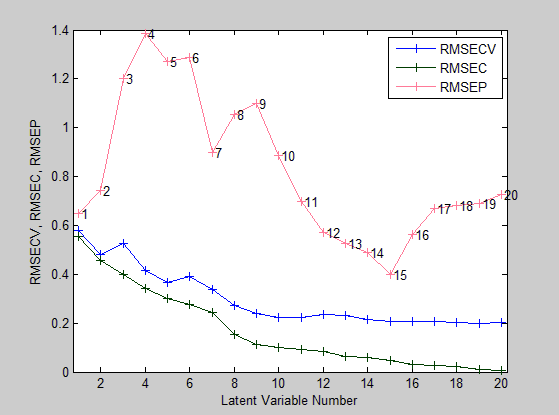
CV Bias: -2.03811e-05

Pred Bias:0.274884

R^2 Cal: 0.95826

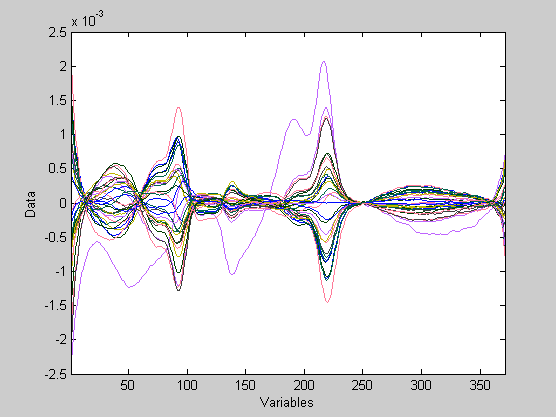
R^2 CV: 0.849594

R^2 Pred: 0.203232



# Combinations

Each of the previous preprocessing methods has been used independently of each other. We will now use a combination of three pre-processing methods to show how powerful of a model we can potentially make. For our combination we will choose a multiplicative scatter correction, followed by a first derivative, and then a mean centering. Doing so results in a rather interesting spectra plot. We see much of the noise has been dealt and it is clear which variables result in the greatest change in spectra.



## SSQ Table

Percent Variance Captured by Regression Model

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

Comp This Total This Total

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

1 79.19 79.19 28.91 28.91

2 15.29 94.48 5.56 34.47

3 2.89 97.37 19.66 54.13

4 1.24 98.61 11.14 65.27

5 0.73 99.34 3.87 69.14

6 0.42 99.76 4.90 74.04

7 0.09 99.85 13.70 87.74

8 0.06 99.91 6.69 94.43

9 0.04 99.95 1.83 96.26

10 0.02 99.96 1.98 98.24

11 0.01 99.97 0.42 98.66

12 0.01 99.98 0.49 99.15

## Prediction

Linear regression model using

Partial Least Squares calculated with the SIMPLS algorithm

Developed 14-Nov-2019 22:00:24.138

Author: kebro@DESKTOP-B4UA9E4

X-block: 15 by 371 (kebro@DESKTOP-B4UA9E4@20191114T153717.93610099 m:20191114154515.881)

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

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

Y-block: 15 by 1 (kebro@DESKTOP-B4UA9E4@20191114T153725.97610212 m:20191114154515.867)

Included: [ 1-15 ] [ 1 ]

Preprocessing: Autoscale

Num. LVs: 12

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

RMSEC: 0.0515509

RMSECV: 0.184059

RMSEP: 0.386416

Bias: -1.77636e-15

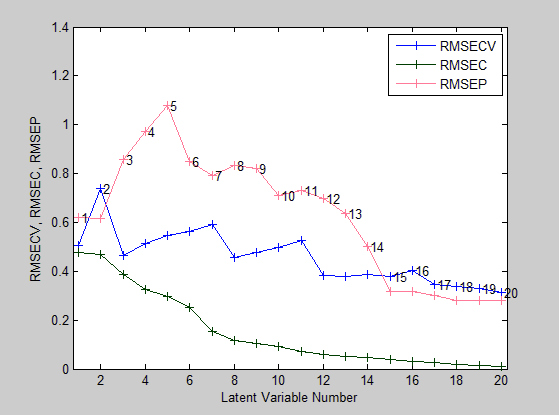
CV Bias: 0.0248788

Pred Bias:0.0678038

R^2 Cal: 0.991504

R^2 CV: 0.909014

R^2 Pred: 0.675853



This pre-processing method resulted in our second-best model. With a cross-validation R^2 of .909 and an RMSE of .184. This model does seemingly generalize to new data well, minimalizing the effect of the present outliers.

# Conclusions

Pre-processing has a more than significant impact in the process of model building. For the purposes of this data set, it appears that applying either a first derivative method by itself or within a combination of MSC and mean-centering results in a powerful model. For the next assignment this will be tested further as unwanted calibration variables and troublesome outliers will be trimmed from the data set and model inputs.