NIR CALIBRATION WITH PCR

A report on Principal Component Regression

Abstract

This report is all about calibrating Near Infrared Spectroscopy obtained values with the machine learning algorithm known as Principal Component Regression, also this case study will prove how PCR is suitable for research purposes in the respective domain and we'll also compare PCR with other machine learning algorithms like PLS and Ridge Regression.

Deep Contractor
deepcontractor@jklu.edu.in

Department of Computer Science and Engineering
Institute of Engineering
JK Lakshmipat University, Jaipur, India

Abstract

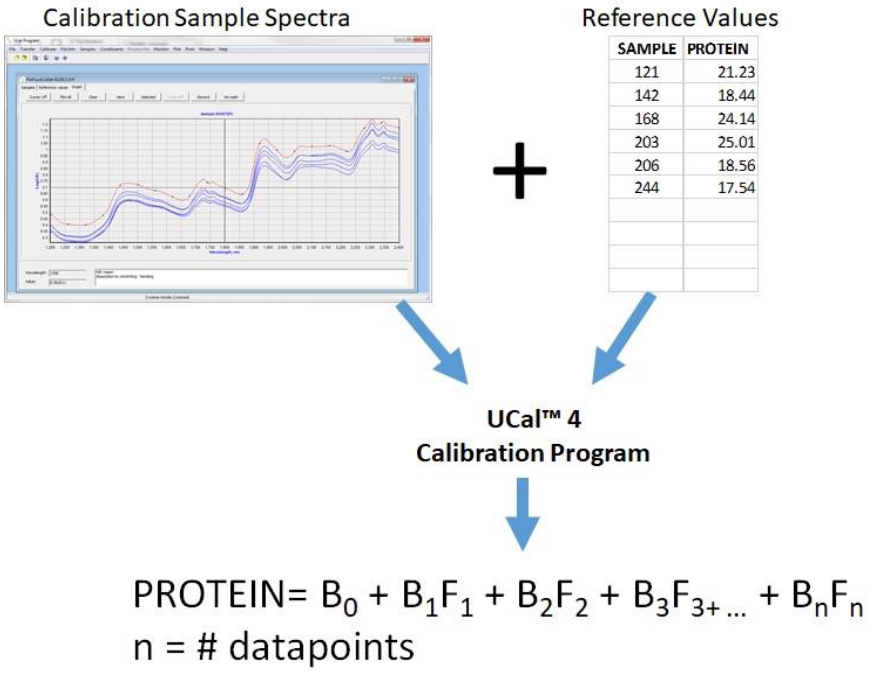
This report is all about calibrating Near Infrared Spectroscopy obtained values with the help of machine learning algorithm known as Principal Component Regression, also this report will help us prove how PCR is suitable for research purposes in the respective domain and we'll also compare the results of PCR with other machine learning algorithms like Partial Least Square and Ridge Regression.

Background

NIR (Near Infrared Spectroscopy) chemometrics is a technique used to measure differences in energies transmitted or deflected from a sample. Here the work calibration is used to point the mathematical co-relation between the NIR data and the chemical property of interest known as brix value which is basically the sugar content of a sample.

The first task to perform NIR calibration is to collect different sample, in our case we have taken Peach (a fruit) which will be used to obtain data points using instrument like FT, PDA am scanning monochromators. To obtain a good result at least 50 samples must be taken. After collecting the samples are sent for reference analysis. Once the data is obtained, they are added to other raw sample and are regressed against each other using different technique such as Principal Component Regression, Partial Least Square and Ridge Regression. The output is a linear equation which is used to predict the brix values of the unknown samples.

Refer to the image attached for a better explanation.



In the above picture you can see how the explained process is executed and result is derived.

Model (PCR)

To calibrate our NIR we'll need to use Principal Component Regression, Now in order to perform the following we'll need to perform the following steps.

1. Divide the data into X and y, Predictor variables and criterion variable.
2. Standardise the Data.
3. Apply PCA on the data to reduce the dimensions.
4. Select the number of principal components that we'll need to use.
5. Perform Linear Regression using the selected Principal Components.
6. Calculate the Model score or the R^2 score

A snapshot of our data:

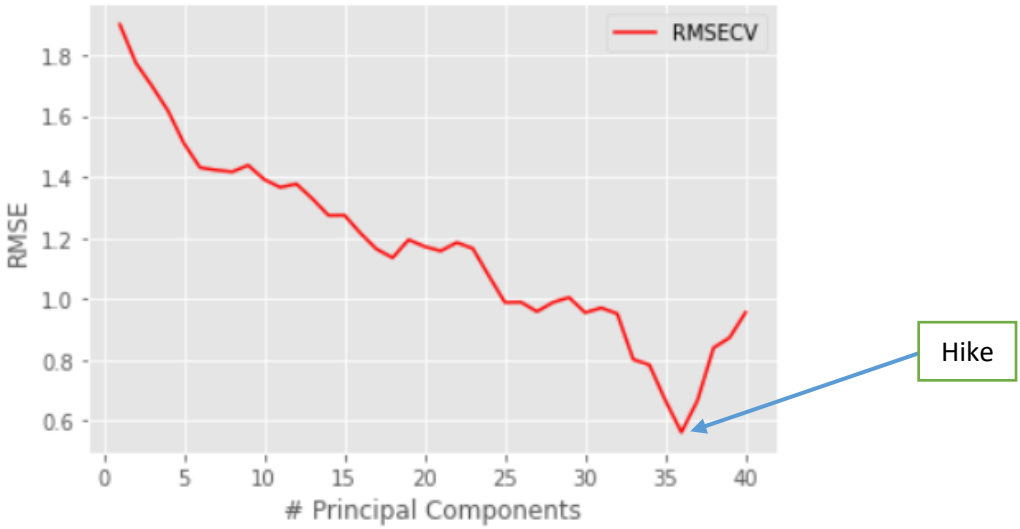
	Brix	wl1	wl2	wl3	wl4	wl5	wl6	wl7	wl8	wl9	...	wl591	wl592	wl593	wl594	wl595	
0	15.5	-1.032355	-1.030551	-1.027970	-1.024937	-1.021866	-1.019143	-1.016866	-1.014910	-1.012907	...	0.692447	0.692944	0.692376	0.690764	0.688081	(
1	16.7	-1.139034	-1.137186	-1.134485	-1.131222	-1.127761	-1.124464	-1.121508	-1.118802	-1.115973	...	0.729328	0.728031	0.725548	0.721815	0.716767	(
2	18.1	-1.152821	-1.150937	-1.148288	-1.145165	-1.141951	-1.138977	-1.136366	-1.134011	-1.131516	...	0.736608	0.735214	0.732669	0.728911	0.723844	(
3	14.8	-1.087215	-1.085455	-1.082867	-1.079797	-1.076568	-1.073632	-1.071087	-1.068877	-1.066654	...	0.758695	0.757963	0.756038	0.752903	0.748496	(
4	15.1	-1.080364	-1.078436	-1.075784	-1.072693	-1.069562	-1.066691	-1.064214	-1.062025	-1.059787	...	0.719793	0.718875	0.716860	0.713771	0.709577	(

Inside variable ‘y’ the first column that is ‘Brix’ is stored as the criterion variable and inside variable ‘X’ we have the data points [column 02-601] also known as the predictor variable.

After assigning the variables to our data now it is time to use sklearn’s StandardScaler to scale our data with the distribution now cantered around 0 with a standard deviation of 1.

With our newly transformed data now it is time to import PCA (Principal Component Analysis) library to perform dimensionality reduction and to scatter our data into different Principal Components.

So the principal components are now derived but now the main problem is to choose the number of principal components that we need to perform regression. There are basically two ways to find the ideal number of principal components. First option is that we can sum up first few of the principal components and use the ones with the most combined variance and our second more appropriate option is to find the RMSEC graph, in my project I have used the second method and obtained the RMSEC graph as shown below:



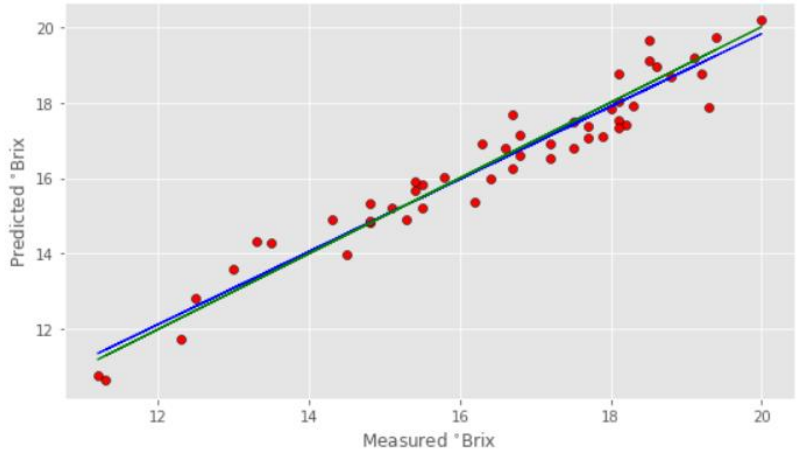
Analysing the above graph it can be noted that at about the x axis value of 37 we can observe a hike which tells us that 37 is the best case scenario for choosing the number of principal components for our regression.

Now that we have decided we move forward to the next step of performing regression with our 37 principle components. After importing sklearn's linear model we perform linear regression on the principle components.

The results of the linear regression were as following:

The score of linear regression model was found to be **0.9250126283699999**

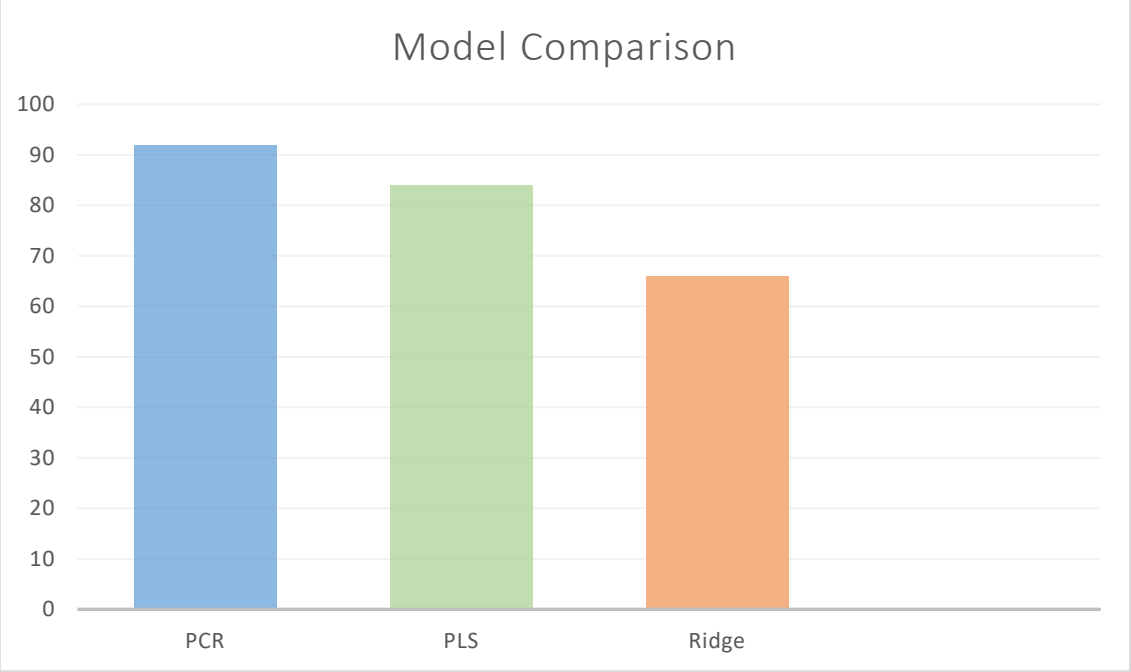
The R^2 score of our regression model was found to be **0.93**



Comparison with other models (PLS and Ridge)

Ridge Regression and Partial Least Squares Regression are some of the best alternatives out there for NIR Calibration, In-fact in some cases PLS outperforms PCR. Let's see how does PLS and Ridge perform in our NIR data when compared to PCR.

Algorithm	No. of PCs	Performance
PCR	37	0.92
PLS	5	0.84
Ridge	-	0.66



In Conclusion the PCR with the most number of Principal Components successfully out performs Ridge Regression and Partial Least Square Regression.

Note : The PLS and Ridge models were optimised to some limit.

Reference

1. https://en.wikipedia.org/wiki/Principal_component_regression#:~:text=In%20statistics%2C%20principal%20component%20regression,principal%20component%20analysis%20(PCA).&text=In%20PCR%2C%20instead%20of%20regressing,variables%20are%20used%20as%20regressors.
2. https://en.wikipedia.org/wiki/Brix
3. https://en.wikipedia.org/wiki/Near-infrared_spectroscopy
4. https://nirpysearch.com/
5. https://learnche.org/pid/latent-variable-modelling/principal-components-regression
6. https://online.stat.psu.edu/stat508/lesson/7/7.1

Appendix A:

A.1 PRINCIPAL COMPONENT REGRESSION:

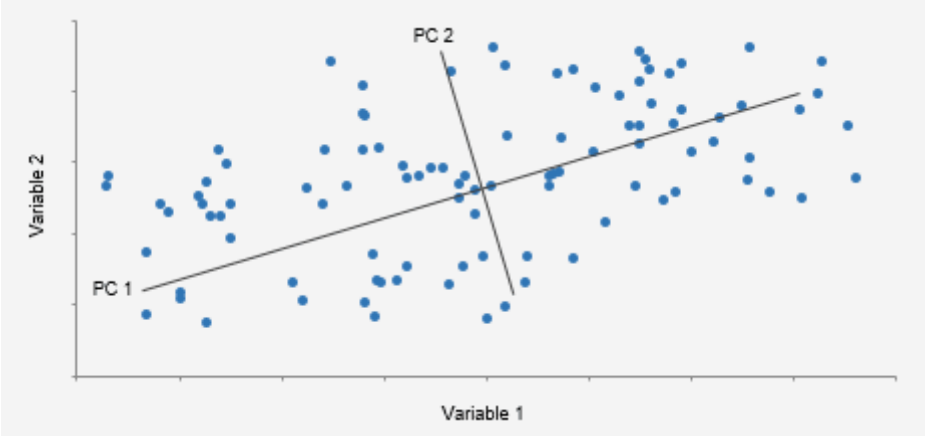
Principal Component Regression (PRC in short) is a sort of an extension to PCA (Principal Component Analysis). To understand PCR we will first need to know what PCA really is?

Principal Component Analysis is an algorithm used to reduce dimensions and supress variations. It can be used to clean data sets and make it easy to explore and analyse. It is totally based on two basic mathematical terms namely :-

1. Variance and covariance
2. Eigen values and Eigen vectors

Principal Components:

The technique of Principal Component Analysis enables us to create and use a set of variables called principal components. As a reduced set of variable is easier and more efficient to analyse and to work on. Suppose you are working on a problem with more than 500 parameters, then PCA will help you reduce it to 5-10 variables with decreasing variance.



PC1 & PC2 are two principal components in the above figure.

Algorithm steps

Step 1: Get your data

Step 2: Give your data a structure

Step 3: Standardize your data

Step 4: Get Covariance of Z

Step 5: Calculate Eigen Vectors and Eigen Values

Step 6: Sort the Eigen Vectors

Step 7: Calculate the new features

Step 8: Drop unimportant features from the new set.

Please note that PCA is NOT a feature selection method, as a feature selection method would involve selecting a few features as it is, out of all of them. So instead, we are combining features to create new PCs, which are different from the original features

These were some steps that we need to follow while performing PCA in our dataset.

Now for Principal Component Regression all we need to do is remove co-related components and apply regression on a set of principal components.

The number of Principle Components that will be used, will always be less than the number of predictors or features that we have. This is because the objective of PCA is to reduce the number of features.

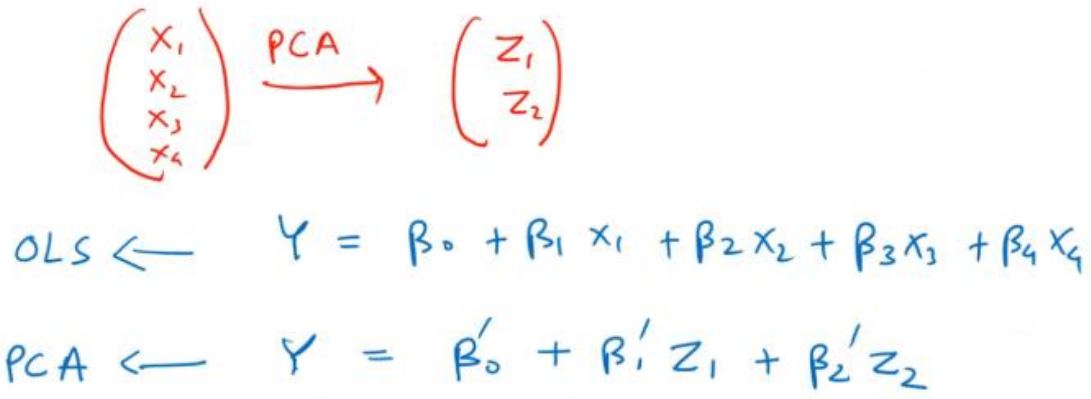
Let's combine all the above steps and make finals steps for PCR.

Algorithm Steps for PCR:

STEP 1: Do PCA to create PCs as our new input features

STEP 2: Use PCs as new input features to train our model for linear regression.

STEP 3: After that we will need to transform those principal components into their original for in order to make prediction on the actual dataset.



That was all the info about the Algorithm, I will conclude with some advantages PCR reduces the number of features of the model

1. PCR is useful on the data with lot of collinearity.
2. It reduces the no. of features.
3. It is useful on dataset with co-related values.
4. It solves the problem of overfitting.
5. Regression is faster because of reduces variables.

Appendix B: CRISP Model Steps:

Appendix B:

CRISP MODEL STEPS:

B.1 Business Understanding

Near-infrared spectroscopy (NIRS) is a spectroscopic technique that uses the near-infrared region of the electromagnetic spectrum (from 780 nm to 2500 nm). Typical applications include medical and physiological diagnostics and research including blood sugar, pulse oximetry, functional neuroimaging, sports medicine, elite sports training, ergonomics, rehabilitation, neonatal research, brain computer interface, urology which is called bladder contraction, and neurology such as neurovascular coupling. There are also applications in other areas as well such as pharmaceutical, food and agrochemical quality control, atmospheric chemistry, combustion research and astronomy.

Device used for NIRS

Brix (symbol °Bx) also known as Degrees Brix is the sugar content of an aqueous solution. One degree Brix is 1 gram of sucrose in 100 grams of solution and represents the strength of the solution as percentage by mass. If the solution contains dissolved solids other than pure sucrose, then the °Bx only approximates the dissolved solid content. The °Bx is used in the wine, sugar, carbonated beverage, fruit juice, maple syrup and honey industries.

B.2 DATA UNDERSTANDING:

The data consists of brix value of 50 peaches (for brix refer business understanding). These individual brix values have around 601 data points making our data High Dimensional Data. Using these data points we can find the brix value of a particular peach fruit.

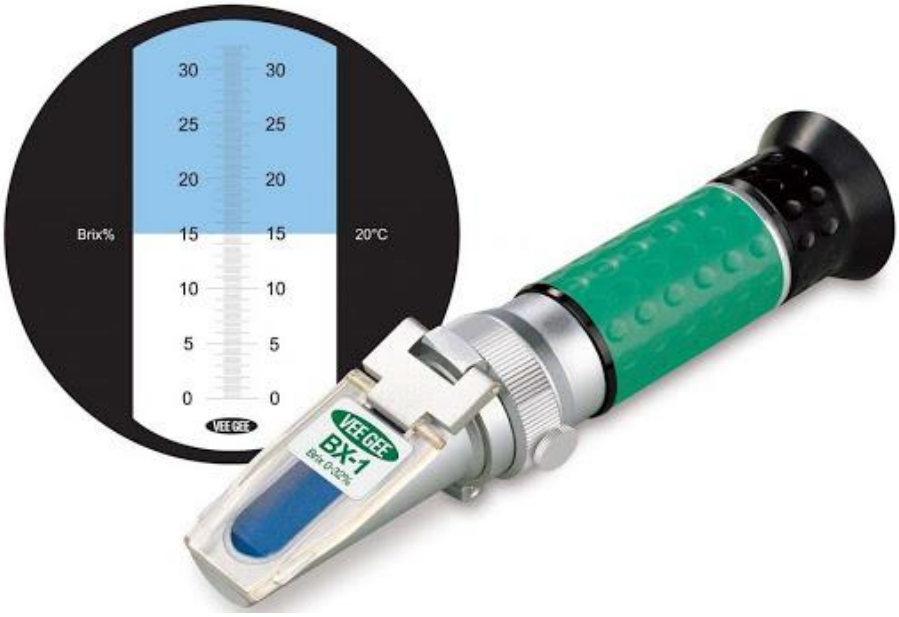
The data consists of around 50 rows and 601 columns of continuous values. The 601 columns are negative values.

B.3 DATA PREPARATION

Data was prepared using various scientific instrument. The brix value of each peaches were obtained using Refractometer.

Also standard scaler is used to standardise the data.

Find below an image of a Refractometer:



B.4 MODELLING

The data is first narrowed down using PCA Principal Component Analysis and using the narrowed down principal components we perform linear regression on them. The model is used to predict the brix values of individual peaches.

The model is also used to find and calculate the score for cross validation and mean square error value of our calibration.

RMSEC and RMSECV values are calculated for individual principal components.

The model is used to find the Coefficient of determination value.

B.5 Evaluation

Using PCR following evaluation were conducted and found. It was found that for conventional PCR the RMSECV value was found higher than the RMSEC value for the individual principal components which shows that the model is a success.

To wrap things up a model score of **0.92** was achieved and R^2 score of 0.93 was achieved.