

A Yokogawa Commitment to Industry

vigilance[™]



Inferential Analysis with NIR and Chemometrics

Santanu Talukdar

Manager, Engineering Services



NIR Spectroscopic Data with Chemometrics A Tutorial Presentation Part 2

References

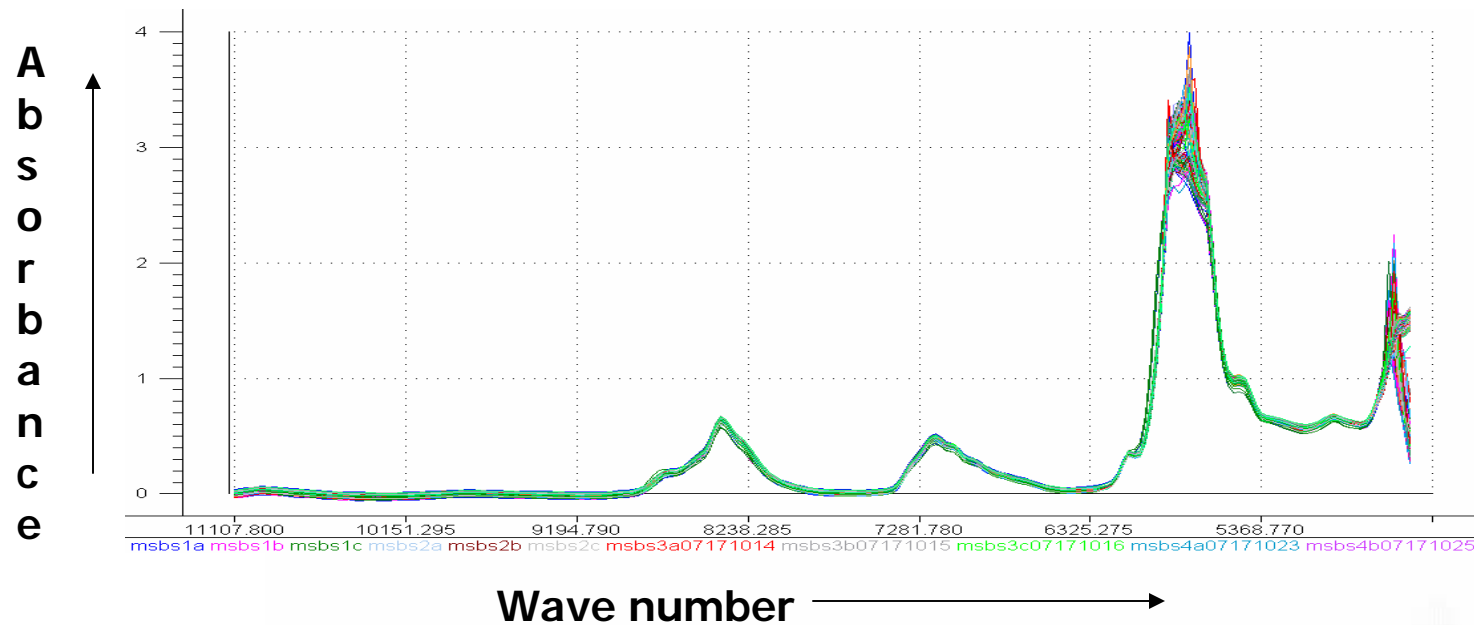
This tutorial is based on the following references:

1. R.De Maesschalck, F.Estienne, J.Verdu-Andres, A.Candolfi, V.Center, F.Despaigne, D.Jouan-Rimbaud, B.Walczak, D.L.Massart, S.de.Jong, O.E.de.Noord, C.Puel, B.M.G.Vandeginste, The Development of Calibration Models for Spectroscopic Data using Principal Component Regression
2. Chemometrics Software, Unscrambler, by CAMO

Yokogawa Corporations India & Japan.



NIR GASOLINE SPECTRUM



File Conversion into *.jdx file

Import of *.jdx into Chemometrics S/W

A Data Matrix

➤ Data Matrix

Wave Nos. →	x1	x2	x3	Xp	Y1	Y2	Y3
S1	a11	a12							
S2									
Samples → S3									
:									
:									
Sm									

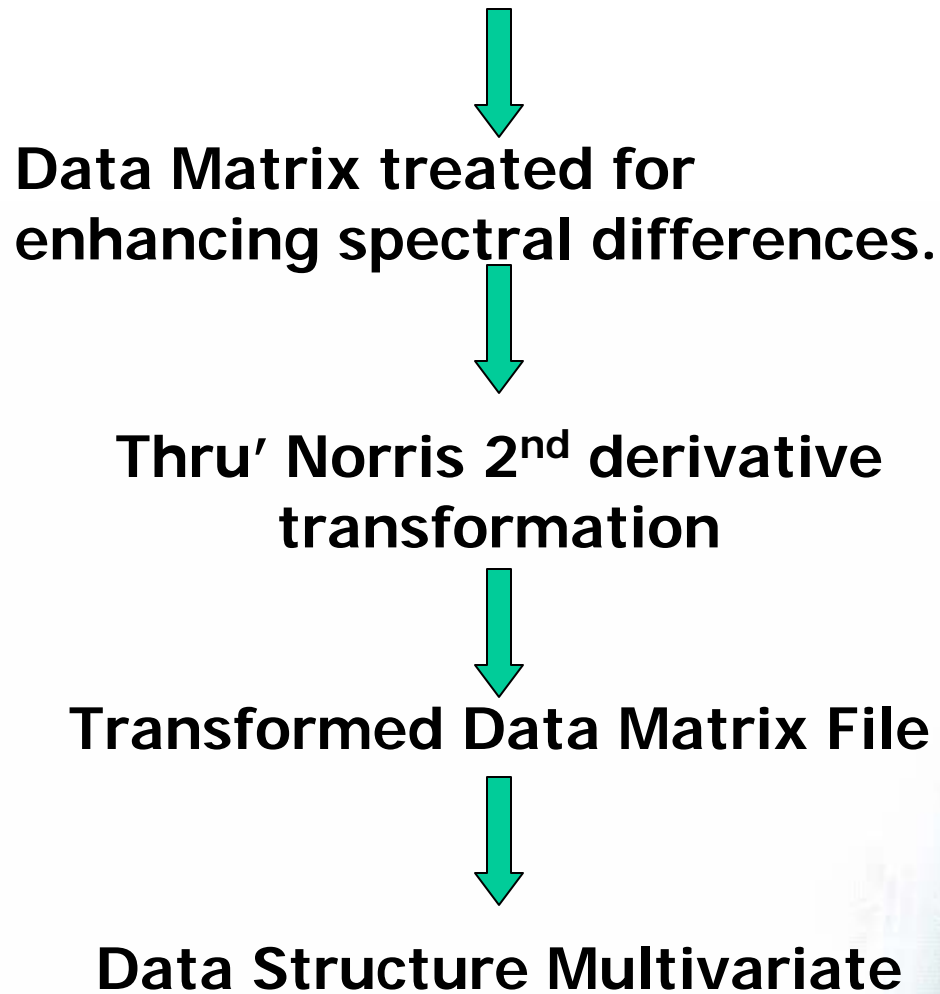
Xm_{xp} Data Matrix

S1...Sm = Samples

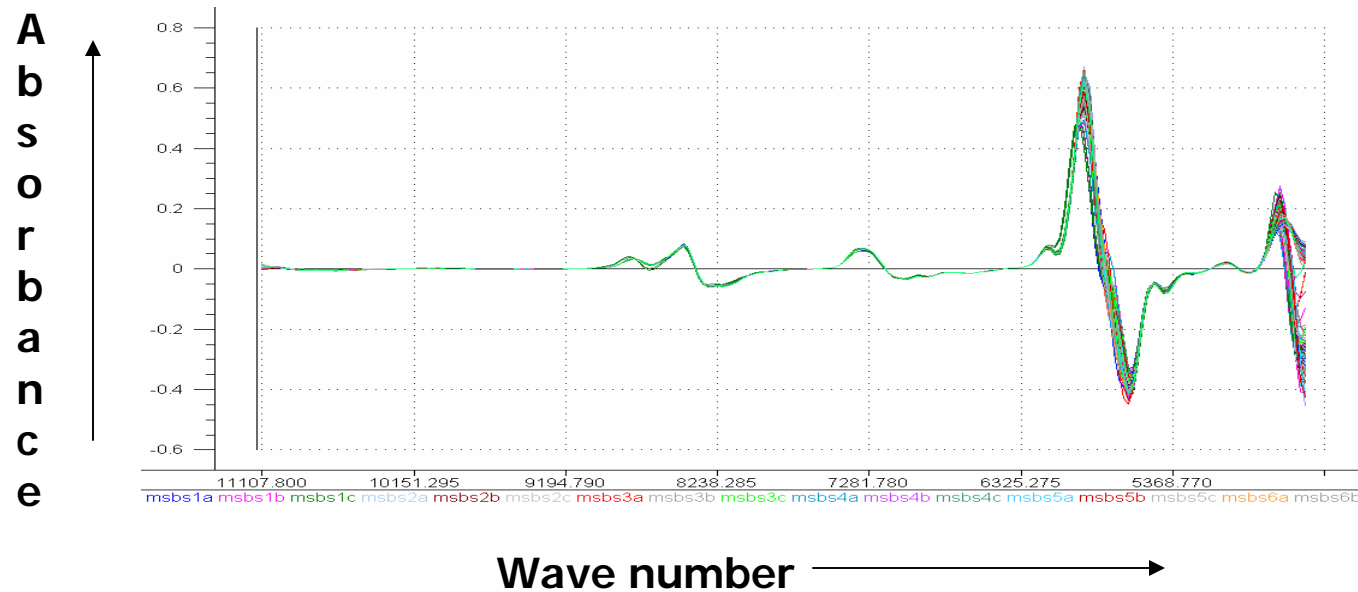
X1...Xp = X variables = Wavenumbers

a_{ij} = Absorbance by NIR Analyzer

Y1...Y3 = Known physical properties (Lab QC) = RON/MON/Etc.

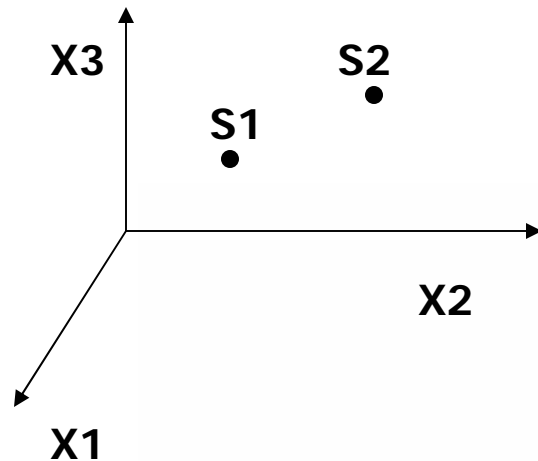


NIR GASOLINE SPECTRUM – NORRIS DERIVATIVE



The change in the Spectrum absorbance after second order Norris derivative transformation.

Representation of a Sample Point in a Vector Space



$$S1 = a11X1 + a12X2 + a13X3$$

$$S2 = a21X1 + a22X2 + a23X3$$

a_{ij} = Cell absorbance

S1 = Sample 1

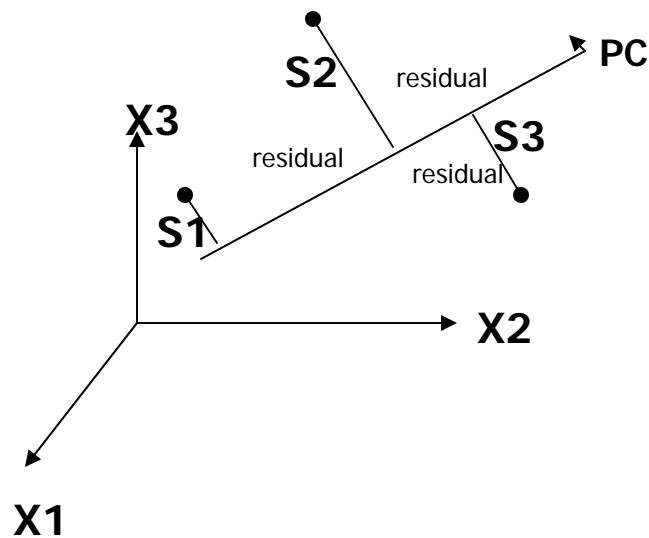
S2 = Sample 2

X1, X2, X3 = The variables

	X1	X2	X3
S1	a11	a12	a13
S2	a21	a22	a23

S1 & S2 are two distinct different samples in 3 dimensional vector space

Principal Component



Location of Sample points along the PC

Principal Component (PC) is a projection in space drawn in such a manner interconnecting sample points and variables such that

- Residual Variances of the samples upon PC is minimum.
- Loading is the cosine of the angle between the PC and each of the variables.
- PC is a linear combination of the variables.
- Number of PCs = Number of Variables.
- All PCs orthogonal to each other.

PC is a projection in a 3 dimensional vector space.

Score Matrix

		PCs							
		PC1	PC2	PC3		PCp	
Samples	S1	t11	t12						
	S2								
	S3								
	:								
	:								
	Sm								

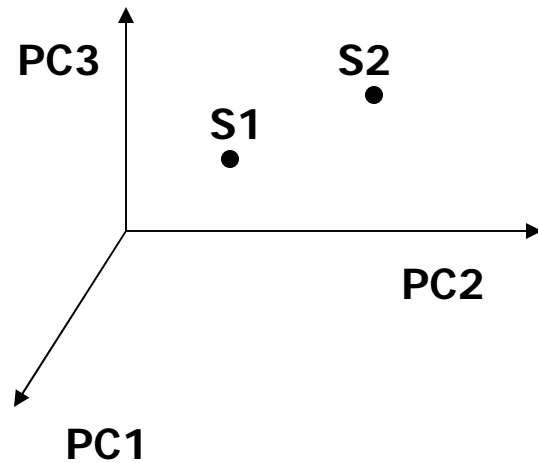
Tm_{xp} Score Matrix

S1...Sm = Samples

PC1...PCp = PCs

t_{ij} = scores

Score Plot in a PC1, PC2 axes



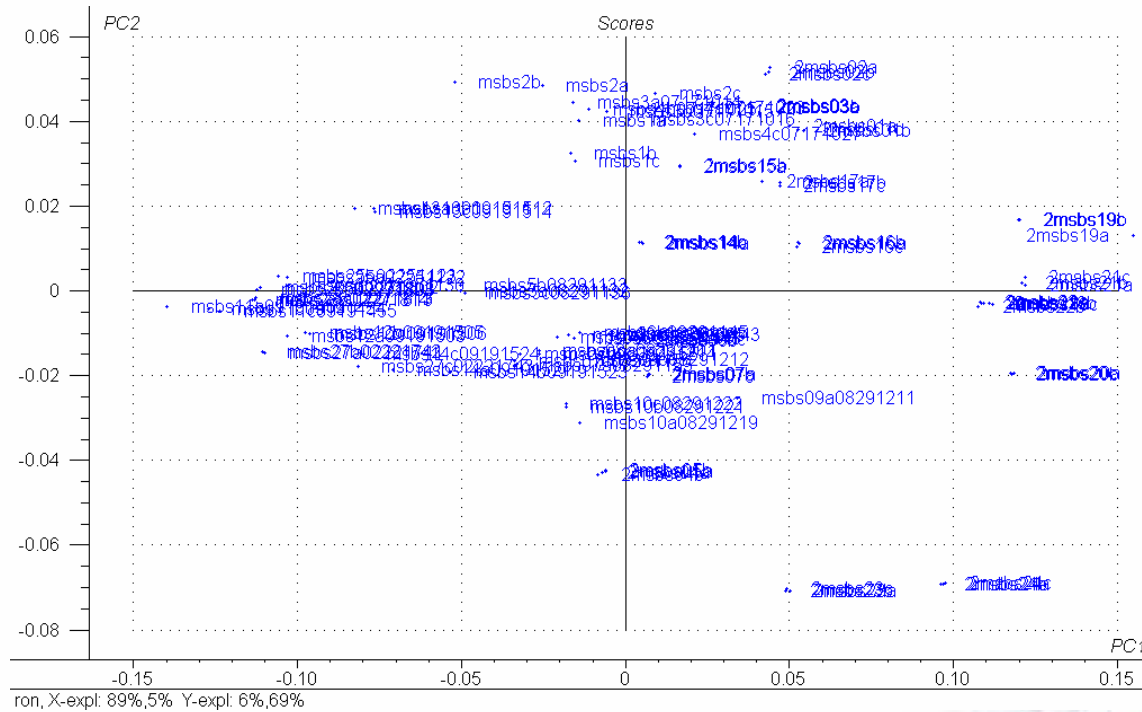
$$S1 = t11PC1 + t12PC2 + t13PC3$$

$$S2 = t21PC1 + t22PC2 + t23PC3$$

	PC1	PC2	PC3
S1	t11	t12	t13
S2	t21	t22	t23

T2x3, a score matrix.

Actual Score Plot in a PC1, PC2 axes



➤ Loading Matrix

PCs.		PC1 PC2 PC3 PCp							
X1	→	p11	p12						p1p
X2									
X3	→								
:									
:									
Xp									

Ppxp Loading Matrix

X1...Xp = Variables

PC1...PCp = PCs

p_{ij} = Loading

→ Transpose of Loading Matrix

	X1	X2	X3				Xp
Variables. →									
	PC1	p11							
	PC2	p12							
	PC3								
	:								
	:								
PCs →	PCp	p1p							

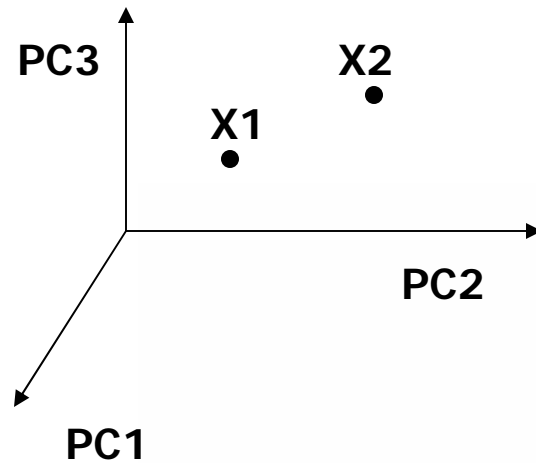
P'p x p transpose of Loading Matrix

X1...Xp = Variables

PC1...PCp = PCs

p_{ij} = Loading

➤ Loading Plot in a PC1, PC2 axes



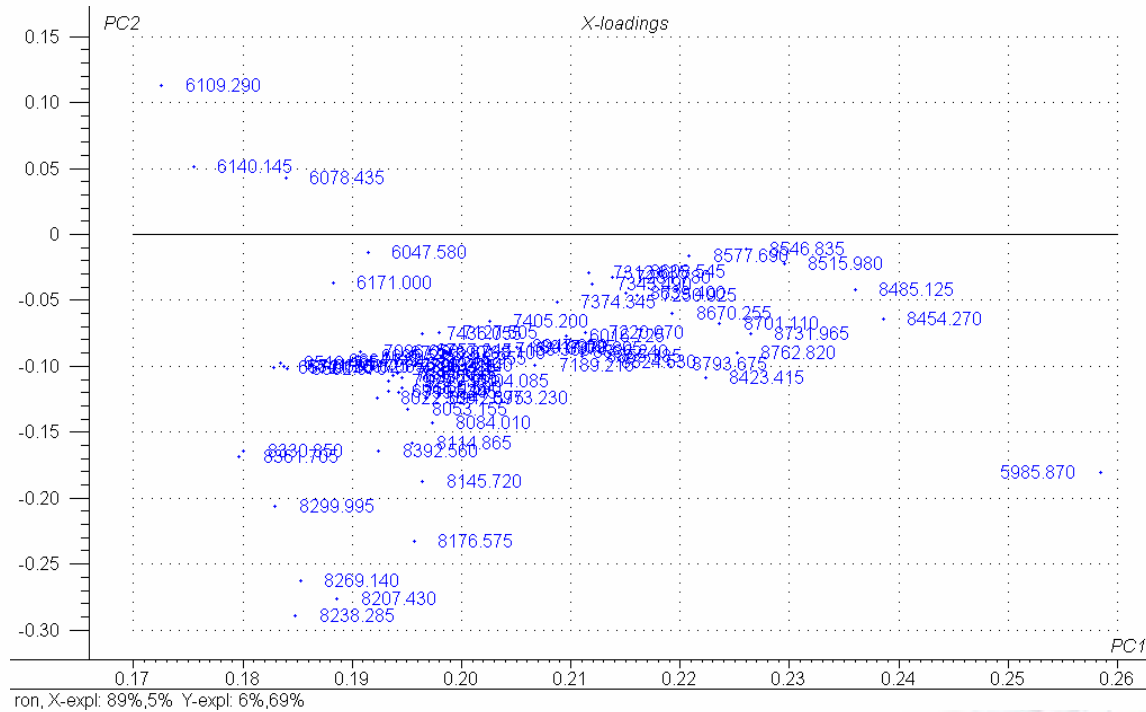
$$X1 = p11PC1 + p12PC2 + p13PC3$$

$$X2 = p21PC1 + p22PC2 + p23PC3$$

	PC1	PC2	PC3
X1	p11	p12	p13
X2	p21	p22	p23

P2x3, a loading matrix.

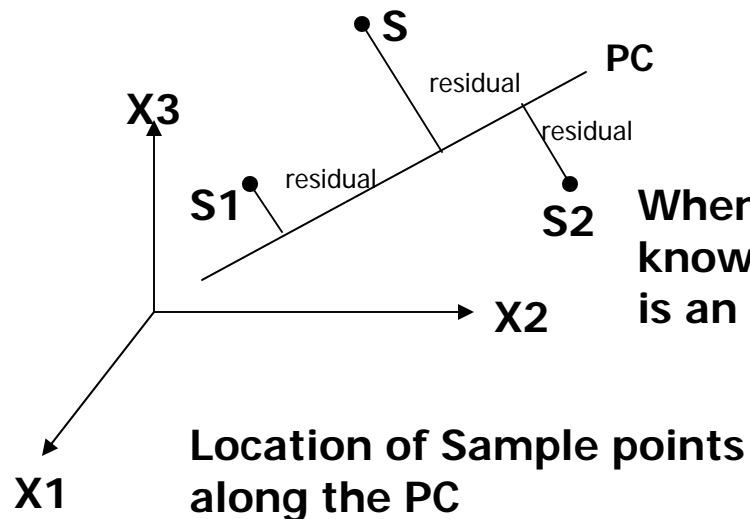
Actual Loading Plot in a PC1, PC2 axes



➤ Residual Variance

Residual Variation of a sample is the sum of squares of its residuals for all the principal components.

It is geometrically interpretable as the squared distance between the original location of the sample and its projection onto the principal component



When any unknown sample S is matched with a known PC & if the residual is high then the sample is an outlier

Principal Component Analysis

$$X_{m \times p} = T_{m \times p} P'_{p \times p} + E$$

$$\text{Data Matrix} = (\text{Score Matrix}) \times (\text{Loading Matrix})'$$

Error is ignored presently.

Principal Component Regression (PCR)

PCR consists of the linear regression of the scores and Y property of interest.

$$\hat{Y} \text{ (predicted)} = T_1 x r \text{ } b r x 1$$

$\hat{Y} \text{ (predicted)} = (\text{row vector with } r \text{ PCs}) \times (\text{column vector with } r \text{ regression coefficients}).$

While doing PCR, only the first r PCs are calculated where $r < \min(m, p)$.

Principal Component Regression (PCR)

How to calculate the column vector b ?

$$Y_{m \times 1} = T_{m \times r} b_{r \times 1}$$

The column vector Y are the known responses from QC Lab.

$T_{m \times r}$ is the score matrix.

Solving the above equation, we get

$$b_{r \times 1} = \text{Inverse}(T'_{r \times m} T_{m \times r}) T'_{r \times m} Y_{m \times 1}$$

Inputting known values of the column vector Y

Column vector b is calculated.

The above equation is known as Multiple Linear Regression (MLR).

The condition is that all variables are linearly independent.

Principal Component Regression (PCR)

Principal Component Regression (PCR) is therefore a two step process.

Step No 1: Decompose the X matrix by PCA.

Step No 2: Fit an MLR model using the PCs instead of raw data as variables.

Hence all PCs are linearly independent.

Variables →

Samples →

	X1	X2	X3				Xp
S1	a11	a12							a1p
S2									
S3									
:									
Sm	am1	am2							amp
Sm'	am'1	am'2							am'p

$X_{m+1 \times p}$ Data Matrix where $X_{m \times p}$ is the old data Matrix and $X_{1 \times p}$ is the new row vector, data appended.

$S_1 \dots S_m$ = Samples, $S_{m'}$ = new sample added to the existing Data Matrix

$PC_1 \dots PC_p$ = PCs

a_{ij} = absorbances by NIR Analyzer

➤ Prediction

For the row vector $S_{m'}$, $a_{m'1}$, $a_{m'2}$, $a_{m'p}$ are the new absorption coefficients generated by the NIR Analyzer.

The new row vector for the new sample $S_{m'}$ can be calculated as

$$T_{1xp} = X_{1xp} P_{pxp}.$$

Select the first r PCs where $r < \min(m,p)$.

$$T_{1xr} = X_{1xp} P_{pxr}.$$

Predicted Y known as

$$\hat{Y} \text{ (predicted)} = T_{1xr} b_{rx1}.$$

❖ Prediction

Also,

$$\hat{Y} = T^T X^T B = X^T P X^T B = X^T B$$

The column vector B are the new regression coefficients.

The X variables are therefore linearly independent.

The response \hat{Y} is therefore directly impacted by the nature of such regression coefficients.

Large positive or negative coefficients will influence the response of \hat{Y} .

In any prediction as the column vector B is constant, large variance is due to the variances in the absorption coefficients in the row vector X^T .

❖ Prediction

Large variances imply samples are of different type.

If extreme samples belong to same population, then a new model has to be prepared or updated to accommodate such extreme samples in the training set.

This implies new regression coefficients generated for the updated model.

Else prepare different models for different sample population.

This problem is known as Sample Clustering and handled by outlier management by proper selection and representation of the calibration sample set.

➤ Prediction Summary

If the new sample is from homogeneous sample population, then the new absorption coefficients are similar from the old data matrix. The error in prediction will be minimum.

If the new sample is from heterogeneous sample population, then the new absorption coefficients will be varying largely from the old data matrix. The new sample will be outlying in the score plot for which the outlying alarm will trigger. The error in prediction will be high.

→ Training Set, Calibration.

Training Set is considered for the initial m number of samples with the Y responses known by prior primary measurements.

This training set is used to develop a model by PCR method. This is known as Calibration.

The regression coefficients are calculated with the help of known Y responses.

Finally, for any unknown sample S , Y is predicted.

➤ RMSEP, Calibration

Each of the samples from the training set are predicted by the developed model.

Each sample is considered as unknown sample for prediction.

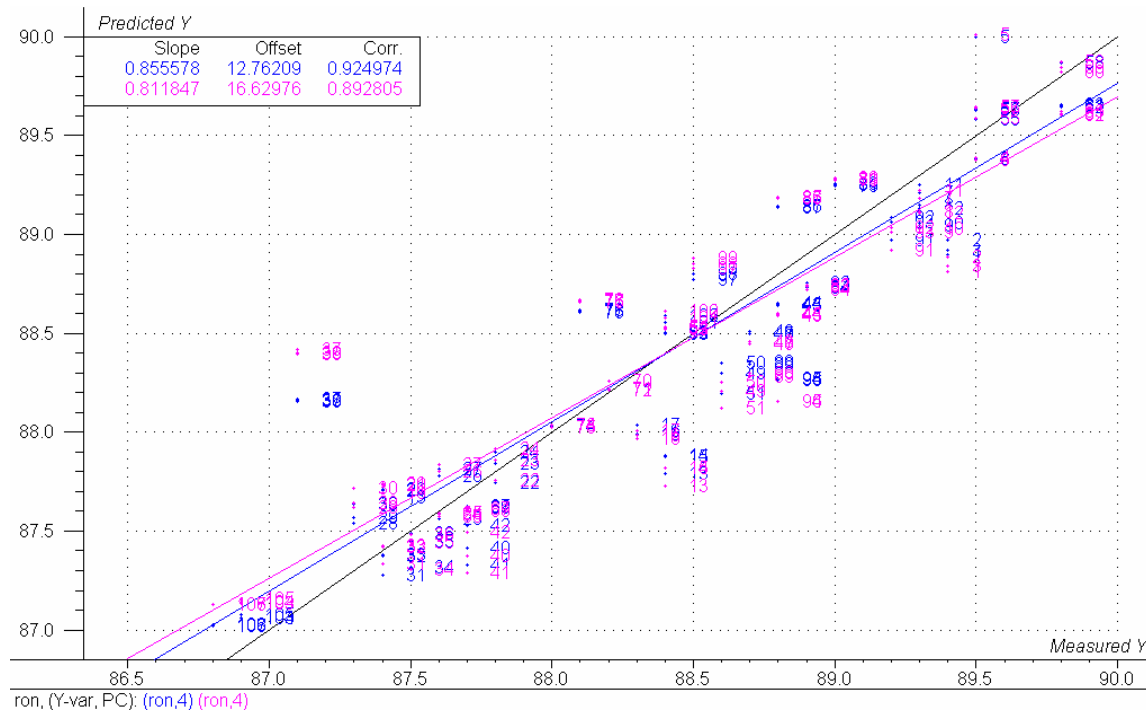
$$e_i = Y_i - \hat{Y}_i$$

Prediction Error Sum of Squares (PRESS) = sum (e_i) squares.

Mean Square Error of Prediction (MSEP) = PRESS /m.

Root Mean Square Error of Prediction (RMSEP) = sq root (MSEP).

➤ Predicted vs Measured Plot



Predicted Y hat for both calibration and validation is plotted against measured Y.

❖ Cross validation.

Data are randomly divided into “d’ cancellation groups.

Suppose there are 15 objects and 3 cancellation groups consisting of objects 1-5, 6-10, 11-15.

The b coefficients in the model that is being evaluated are determined first for the training set consisting of objects 6-15 and objects 1-5 function as a test set, i.e they are predicted with this model.

Then a model is made with objects 1-5 and 11-15 as training set and 6-10 as test set.

Finally a model is made with objects 1-10 in the training set and 11-15 in the test set.

➤ RMSEP, Validation

PRESS is determined for each of the “d” cancellation groups.

Eventually the d PRESS is added to give a final PRESS.

RMSEP is calculated from the final PRESS.

→ Optimization.

Optimization consists of comparing different models and deciding which one gives best prediction.

In PCR, the usual procedure is to determine the predictive power of models with 1, 2, 3...PCs and to retain the best one.

For Example

Select the first PC. Perform PCR & calculate RMSEP1.

Select 1 & 2 PCs. Perform PCR & calculate RMSEP2.

::

Select 1, 2, 3,r PCs. Perform PCR & calculate RMSEPr.

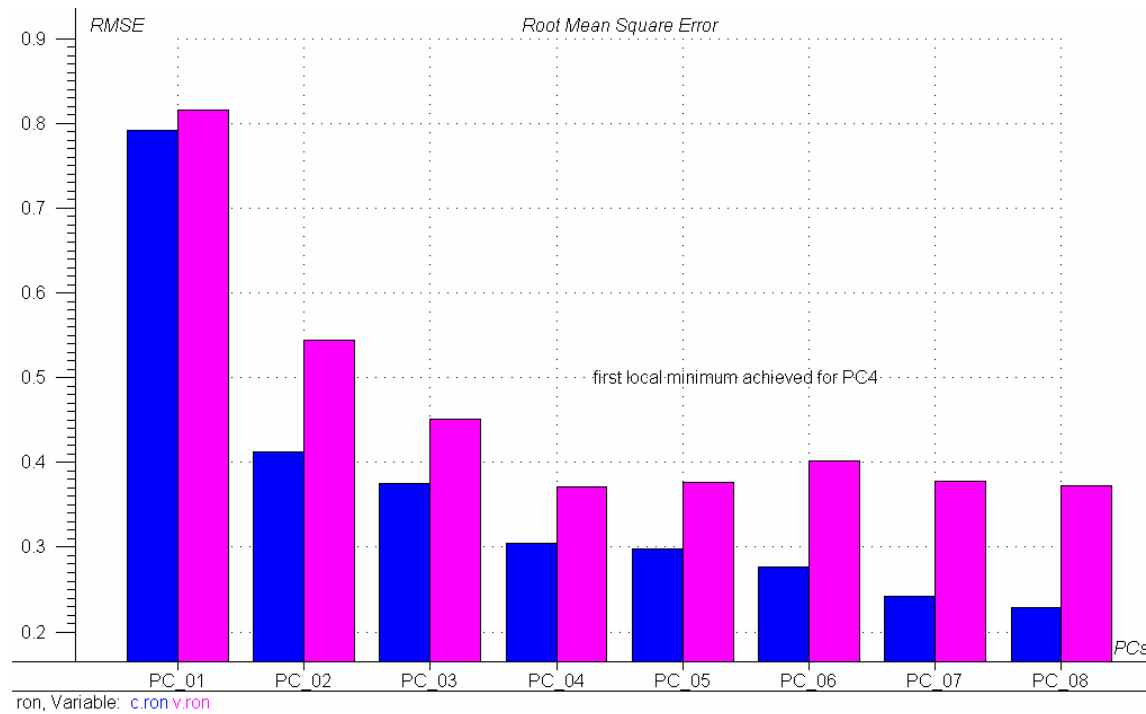
➤ Optimization.

The result is presented as a plot showing RMSEP as a function of the number of components and is called the RMSEP curve.

This curve often shows an intermediate minimum, the first local minimum or a deflection point and the number of PCs for which this occurs is then considered to be the optimal complexity of the model.

The robust model uses this local minimum or the first deflection point rather than the global minimum.

RMSE Plot



External Validation.

External Validation uses a completely different group of samples for prediction (called the test set) from the one used for building the model (the training set).

Both the sample sets are obtained in such a way that they are represented for the data being investigated.

With an external test set the prediction error obtained may depend to a large extent on how exactly the objects are situated in space in relation to each other.

Uncertainty in prediction error can be represented as "Prediction +/- 2*RMSEP".

This measure is valid provided that the new samples are similar to ones used for calibration, otherwise the prediction error might be much higher.

❖ Selection and Representation of the Calibration Sample Set

All possible sources of variation that can be encountered must be included in the calibration set.

Sources of variation such as of different origins or different batches are included and possible physical variations (e.g different temperatures, pressures, flow, etc) among samples are also covered.

One approach for selecting representative calibration is the possibility based on knowledge about the process operational changes.

Another approach is based on D-optimal concept.

❖ Selection and Representation of the Calibration Sample Set

The D-optimal criteria minimizes the variance of the regression coefficients. This is equivalent to selecting samples such that the variance is maximized. Variance maximization leads to selection of samples with relatively extreme characteristics and located on the borders of the calibration domain.

➤ Example

Let the two points as response variable Y selected be 87, 89.

The two new candidate points for selection be 87.5 and 90.

First select the minimum distances of the new candidate points, from the selected points as $d1 = (87.5 - 87 = 0.5)$, $d2 = (90 - 87 = 3)$, $d3 = (89 - 87.5 = 1.5)$, $d4 = (90 - 89 = 1)$.

Selected distances are $d1$ and $d4$.

Select the maximum of $d1$ and $d4$.

$d4$ is the new selected point.

The selected samples are now 87, 89, 90.

Quantitative Evaluation

Y_i = Set of known response variables.

Each response variable for point i is measured 3 times for the total span.

Average Standard Deviation SD of each variable is calculated.

Historical SD of variable = $2.7 * \text{Avg SD}$.

Repeatability = Historical SD.

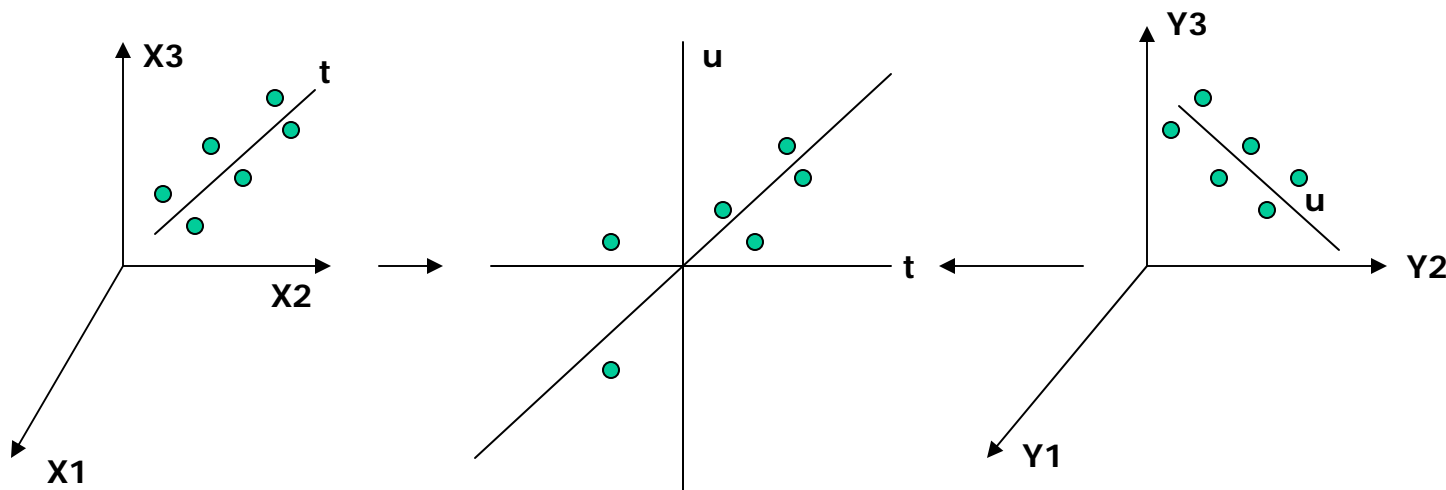
Select samples such that the minimum range of $Y = 5 * \text{repeatability}$ but not less than $3 * \text{repeatability}$.

Reference: ASTM

PLS : Partial Least Squares

Partial Least Squares or Projection to latent structure. Models both the X & Y matrices simultaneously to find the latent variables in x that will predict the latent variables in Y the best.

These PLS-Components are similar to principal components and will also be referred to as PCs.



$$PCy = f(PCx)$$

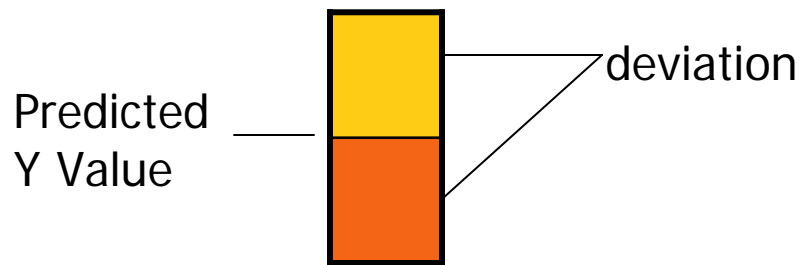
$$u = f(t)$$

PLS1 deals with only one variable at a time (like PCR)

PLS2 handles several responses simultaneously

❖ Prediction Outlier

Predicted Value Vs Reference Value (QC Lab)



This is a plot of predicted Y-Value for all prediction samples. Boxes around the predicted value indicate the deviation.

A large deviation indicates that the sample used for prediction is not similar to the samples used to make the calibration model

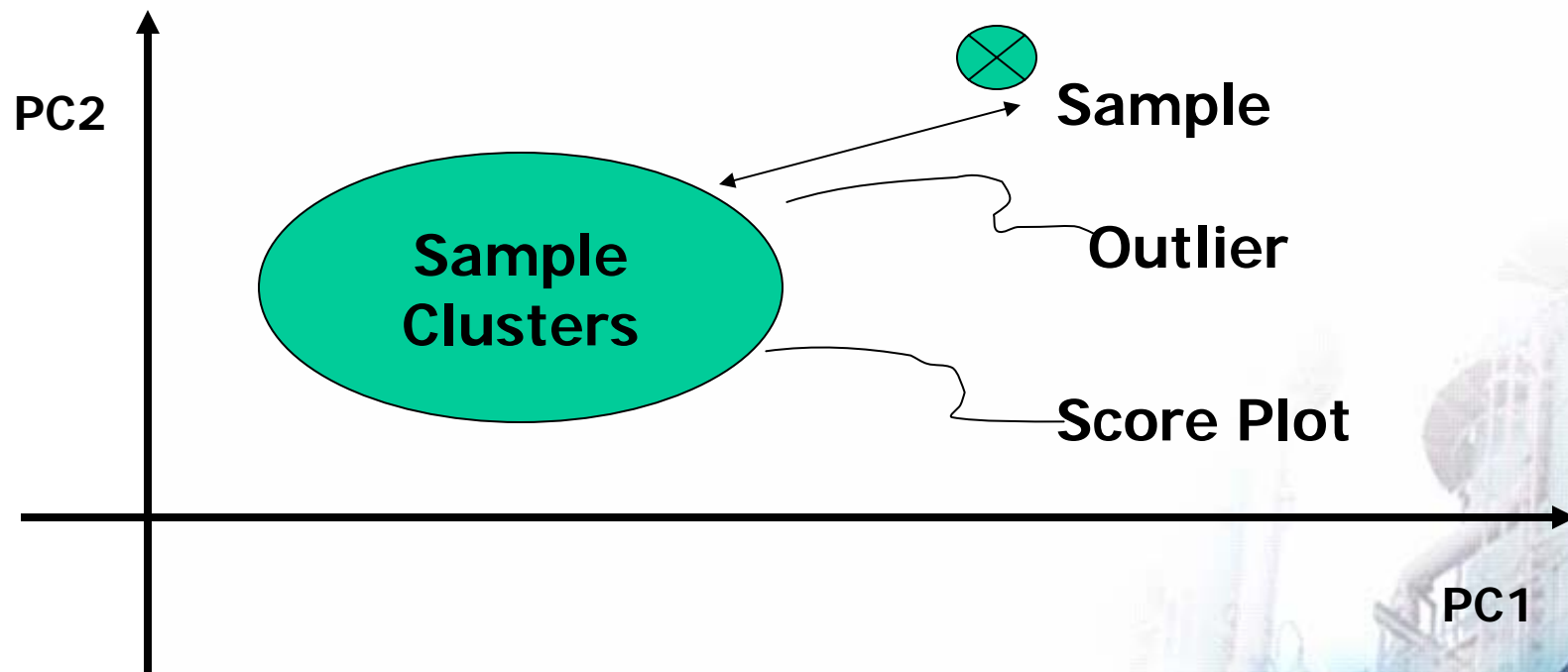
This is a prediction outlier.

Conclusion is that the prediction sample does not belong to the same sample population as the samples the model is based upon.

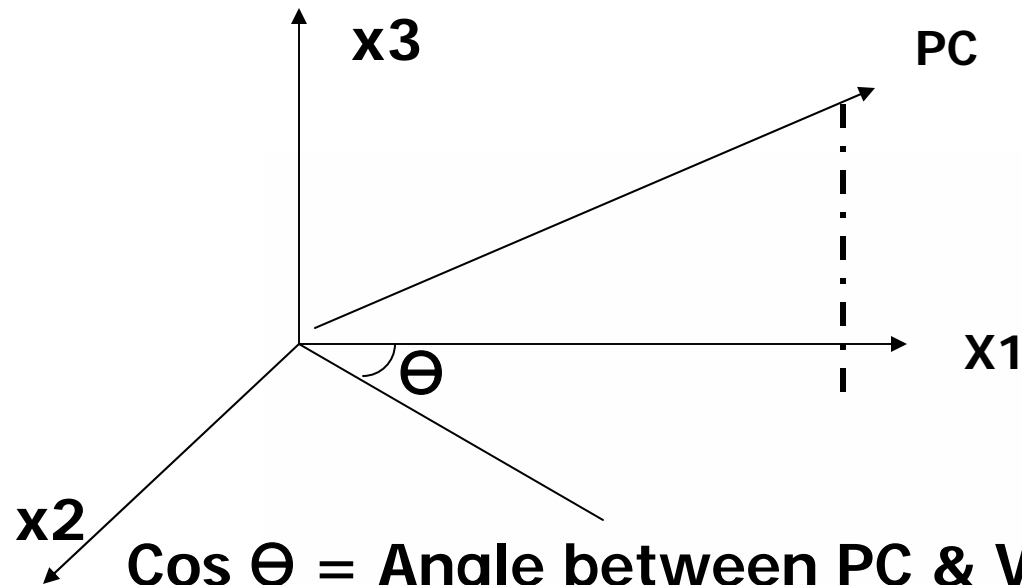
➤ Outlier in Score Plot

Score is the co-ordinate of sample along the PC axes.

The outlying sample is clearly distinct from the sample cluster.



➤ Loading



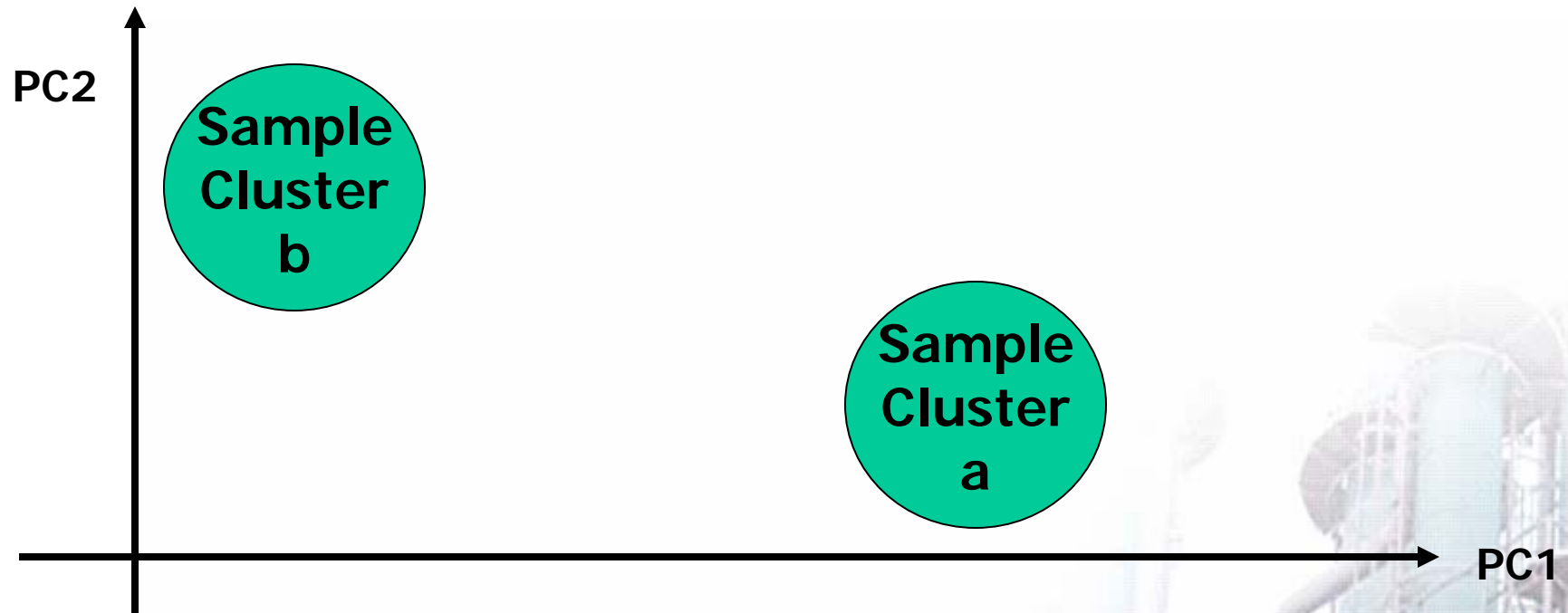
$\cos \theta = \text{Angle between PC \& Variable } X_i$
= Loading

$$-1 \leq \text{loading} \leq +1$$

Two Variables having +values are in +ve correlation

Two Variables having +values & -values are in –ve correlation

Cluster in Score Plot



Sample Clusters {Different Sample clusters a & b due to change in sample population type, although variables are +ve.}

➤ Clusters in Data Structure

Different sample clusters lead to more inaccuracy in models

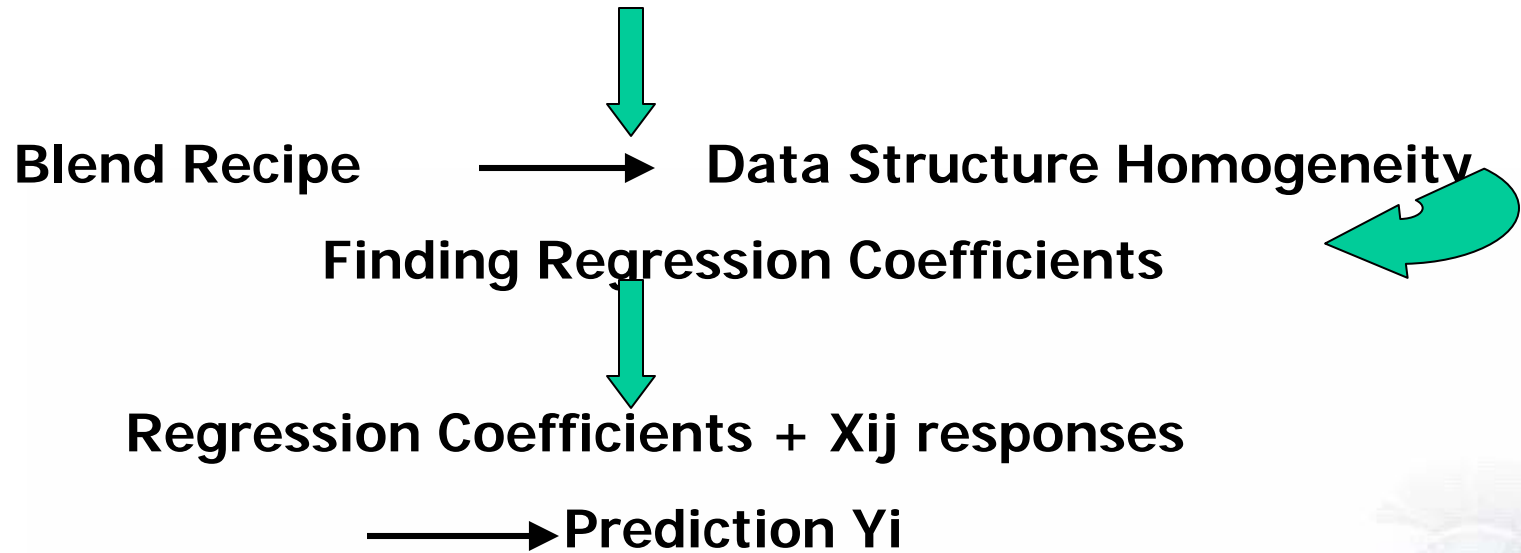
**This can happen due to different recipes
{Different sample Population}**

$$Y_i = \sum B_{ij} X_{ij} + \text{Constant}$$

In a data structure, with known Y_i (Lab values with known variables X_{ij} , coefficients B_{ij} is found out).

In unknown case with similar data structure, Y_i is predicted depending upon X_{ij} responses

B_{ij} are known as Regression Coefficients



Hence,

Blend Recipe Change → Different Sample Clusters
→ Cause Outliers → Change in Regression coefficients

Thank you very much for your attention.

Comments : ?

E-mail :

Santanu.Talukdar@in.yokogawa.com

A Yokogawa Commitment to Industry

vigilance™