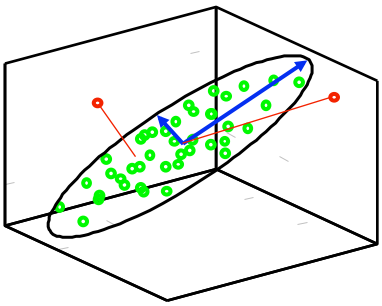
Chemometrics

Without Equations
(or hardly any)



Contact Information

Don Dahlberg
Department of Chemistry
Lebanon Valley College
101 North College Avenue
Annville, PA 17003-1400
Ph: (717) 838-1269
dahlberg@lvc.edu

Barry M. Wise
Eigenvector Research, Inc.
196 Hyacinth Road
Manson, WA 98831
Ph: (509) 662-9213
bmw@eigenvector.com
www.eigenvector.com

Biometrics

Technometrics

Econometrics

Psychometrics

Chemometrics

Chemical

Measurements

While worrying about the "Metrics" part, do not forget the "Chemo" part.

3

EIGENVECTOR RESEARCH INCORPORATED

Chemometrics - Use of Mathematics, Chemistry, and Logic to Perform:

- Experimental Design - How to take measurements in such a way as to maximize the chances of obtaining the desired information at the least cost.
- Data Analysis - How to get as much of the information out of a set of measurements as possible and relate *measurements* made on a *chemical* system to the *state* of the system

4

EIGENVECTOR RESEARCH INCORPORATED

This Workshop Concentrates on Two Aspects of Data Analysis:

- Exploratory Data Analysis and Pattern Recognition
- Regression

5

EIGENVECTOR RESEARCH INCORPORATED

Mathematical Tools Used in Chemometrics Can be Very Sophisticated.

- Multivariate Statistics
- Matrix and Tensor Algebra
- Eigenfunction - Eigenvalue Problems
- Neural Networks, Support Vector Machines
- Discriminant Analysis
- Etc.

But the Software Can Take Care of the Calculations -

Provided We Understand Conceptually What the Software is Doing.

6

EIGENVECTOR RESEARCH INCORPORATED

Outline (Part 1)

- Introduction
- Exploratory Data Analysis & Pattern Recognition Motivation
- Principal Components Analysis
- SIMCA
- Summary

7

EIGENVECTOR RESEARCH INCORPORATED

Outline (Part 2)

- Regression Motivation & Rational
- Classical Least Squares (CLS)
- Multiple Linear Regression (MLR)
- Principal Components Regression (PCR)
- Partial Least Squares Regression (PLS)
- Summary

8

EIGENVECTOR RESEARCH INCORPORATED

Outline (Part 1)

- Introduction
- Motivation for Exploratory Data Analysis & Pattern Recognition
  - what is Exploratory Data Analysis?
  - what is Pattern Recognition?
  - relevant measurements
  - some statistics definitions
- Principal Components Analysis
- SIMCA
- Summary

EIGENVECTOR RESEARCH INCORPORATED

What is Exploratory Data Analysis and Pattern Recognition?

It is important to examine data patterns to:

- Identify trends, clusters, and other patterns that should be explained and understood before modeling.
- Ensure the data represents all types of samples that the model will be expected to handle in application.
- Detect and correct or remove faulty samples which might endanger the reliability of the model when applied to future samples.
- Often the final goal is to identify trends or class membership.

EIGENVECTOR RESEARCH INCORPORATED

Pattern Recognition Applications

- How do I know I am producing a product within or outside of specification(s)?
- Am I producing the same product at all my plants?
- What new source of a feed-stock is most like the source I just lost?
- How should I modify my formulation to correct for a new feed-stock to produce product closest to my traditional product?

These are just different ways of asking the same question:

How do if I know when two items are the same or when they are significantly different?

To Answer Any Such Questions I Need Information - Data

Data is usually easy to find:

- Quality Monitoring
- Product Research
- Customer Surveys
- etc.

Data from designed experiments is usually best

13

EIGENVECTOR RESEARCH INCORPORATED

Chemical Instrumentation Has Caused a Data Explosion

FTIR Spectra
60 samples from
3500-700 cm^{-1} @ 4 cm^{-1} resolution

84,000 Pieces of Information!

Vis-NIR-XRF Hyperspectral image

606 x 659 pixels = 399,354 samples
from 1241 channels / wavelengths

495,598,314 Pieces of Information!

14

EIGENVECTOR RESEARCH INCORPORATED

Most of This Information is Either:

- Irrelevant to the problem we want to solve
- Redundant
  - redundancy is a problem for some types of analyses
  - opportunity for other types

Need to find a way to sift through this data and find the parts that will lead to a solution to the problem

Trivial Example: How can one tell the difference between a cat and a dog?

Perhaps Count the Number of Feet?

Fluffy	4
Spot	4
Puff	4

Rex	4
Lassie	4
Rover	4

These numbers are all the same.

No Variance in this Variable

Therefore no information in these numbers.

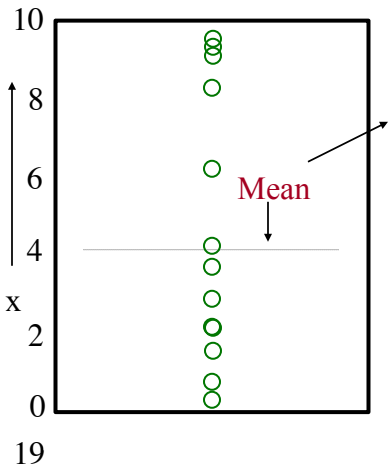
EIGENVECTOR RESEARCH INCORPORATED

A number only has value when it is compared with other numbers

Since we are looking for the difference between dogs and cats, we need a numerical descriptor that changes, *i.e.* has variance

EIGENVECTOR RESEARCH INCORPORATED

Before Defining Variance
Need to Define Mean:



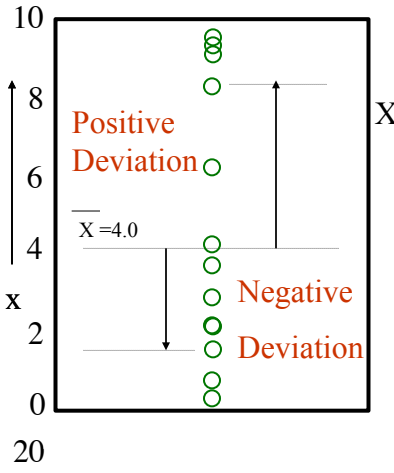
Bar signifies the mean value

$$\overline{X} = \frac{9.4 + 4.1 + 0.6 + 3.5 + \ldots + 2.7}{15} = 4.0$$

- Degrees of Freedom
- Number of samples
- Number of pieces of information available

EIGENVECTOR RESEARCH INCORPORATED

Calculating Variance



Deviations squared to make positive

$$X_{Var} = \frac{(9.4-4.0)^2 + (4.1-4.0)^2 + \ldots + (2.7-4.0)^2}{15 - 1} = 11.7$$

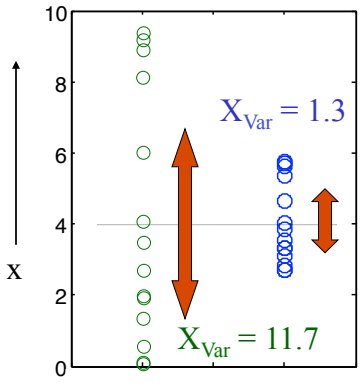
- Lost one Degrees of Freedom

Why?

Conservation of Information

EIGENVECTOR RESEARCH INCORPORATED

Both data sets have the same mean



but they have different variance

Note: Standard Deviation

$$X_{std} = (X_{var})^{1/2}$$

EIGENVECTOR RESEARCH INCORPORATED

But in statistics ...

Variance = Information

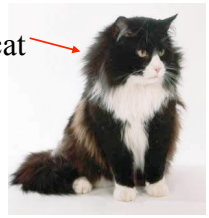
Note: Many methods are concerned with capturing the maximum sum of squares or with minimizing the residual sum of squares (*i.e.* least squares)

Back to the Cats and Dog problem

EIGENVECTOR RESEARCH INCORPORATED

Perhaps length of fur can be used to tell cats from dogs?

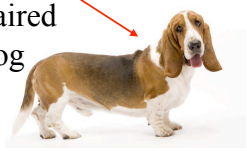
Long haired cat



Short hair cat

Large Variance

Short haired dog



Long haired dog

But still little Discrimination Power between cats and dogs.

23

EIGENVECTOR RESEARCH INCORPORATED

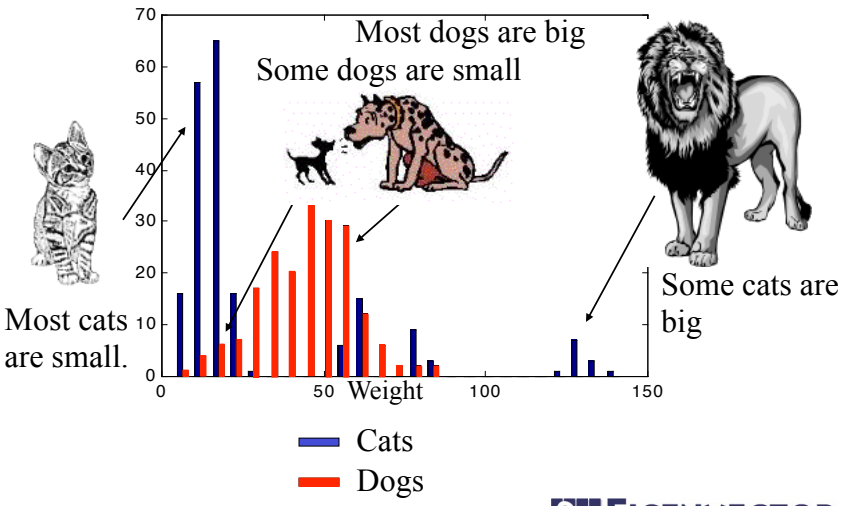
Variance = Information

But is the information relevant to the problem or stated objective?

24

EIGENVECTOR RESEARCH INCORPORATED

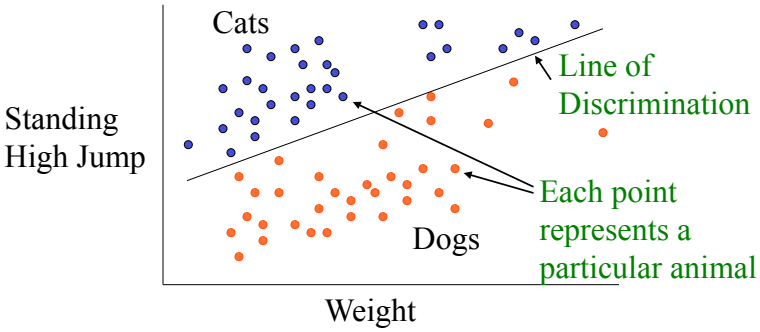
Perhaps Weight of Animal?



25

EIGENVECTOR RESEARCH INCORPORATED

Perhaps two variables *Together*?



26

EIGENVECTOR RESEARCH INCORPORATED

Some Questions:

- If two variables were better than one, how about three or more?
- How to determine which variables are relevant to the problem?
- How to determine the underlying variables that define the problem and the solution?
  - I want to understand the process (*i.e.*, the underlying structure of the data).

A more realistic example…

EIGENVECTOR RESEARCH INCORPORATED

Table of 20 Solutions Analyzed for Ca^{2+} and Cl^- Concentrations.

	$[Ca^{2+}]$	$[Cl^-]$
1	0.1254	0.2435
2	0.4542	0.9097
3	0.7190	1.4361
4	0.8965	1.7946
5	0.2770	0.5536
⋮	⋮	⋮
20	0.9944	1.9886

20 Rows

Samples

20 x 2 Matrix

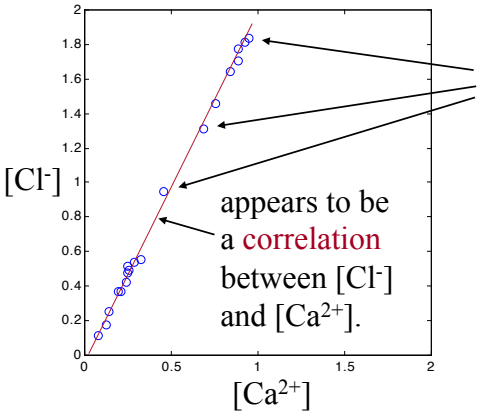
table of numbers

2 Columns

Variables

EIGENVECTOR RESEARCH INCORPORATED

Plot [Cl^-] versus [Ca^{2+}]



Each point represents a sample in two-variable space.

The existence of strong correlation implies there aren't two variables, but two forms of the same underlying variable.

EIGENVECTOR RESEARCH INCORPORATED

How to find this underlying variable?

Next topic:

Principal Component Analysis

PCA

EIGENVECTOR RESEARCH INCORPORATED

Outline (Part 1)

- Introduction
- Pattern Recognition Motivation
- Principal Components Analysis (PCA)
  - What is PCA?
  - Scores and Loadings
  - Interpretation
  - Examples
- SIMCA
- Summary

EIGENVECTOR RESEARCH INCORPORATED

The Problem in Dealing with So Many Variables is that I Can Only Think in 2 or 3 Dimensions

Is there a way to concentrate the information from *many* variables into a *few* (hopefully 2 or 3) underlying variables?

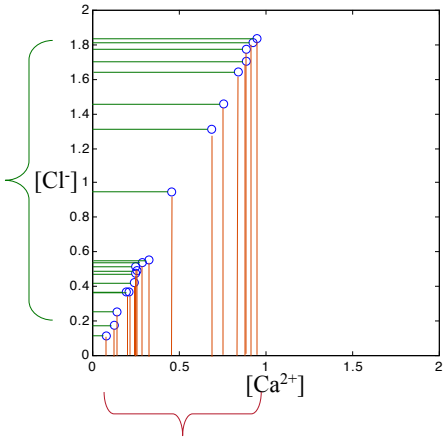
Hint: Variance = Information

EIGENVECTOR RESEARCH INCORPORATED

Plot the Cl^- Concentration vs. the Ca^{2+} Concentration

$[Cl^-]_{var} = 0.4557$

Can do the same with $[Cl^-]$.



$[Ca^{2+}]_{var} = 0.1135$

Can read the $[Ca^{2+}]$ of each sample by projecting it on the [Ca2+] axis.

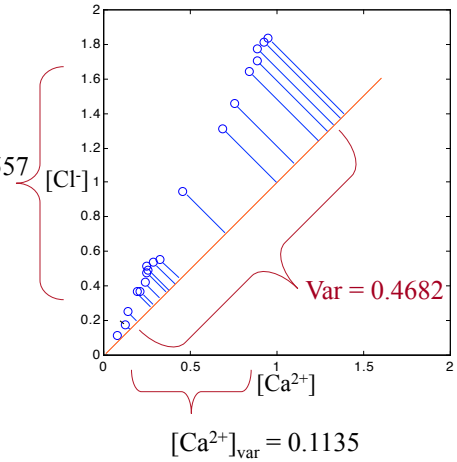
Can then calculate the variance of the [Ca2+].

33

EIGENVECTOR RESEARCH INCORPORATED

Find a New Axis of Maximum Variance:

$[Cl^-]_{var} = 0.4557$



$[Ca^{2+}]_{var} = 0.1135$

There are an infinite number of new axes that I can create.

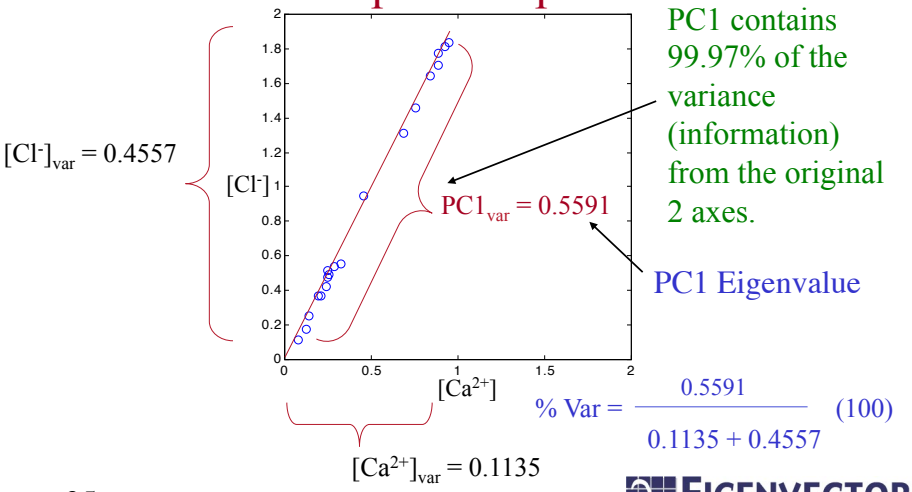
I can project my samples onto this new axis and calculate the amount of information captured by this axis (i.e. variance.)

34

EIGENVECTOR RESEARCH INCORPORATED

Only One New Axis Which Captures the Maximum Variance:

Principal Component



PC1 contains 99.97% of the variance (information) from the original 2 axes.

PC1 Eigenvalue

$$\% \text{ Var} = \frac{0.5591}{0.1135 + 0.4557}(100)$$

35

EIGENVECTOR RESEARCH INCORPORATED

How Was PC1 Constructed?

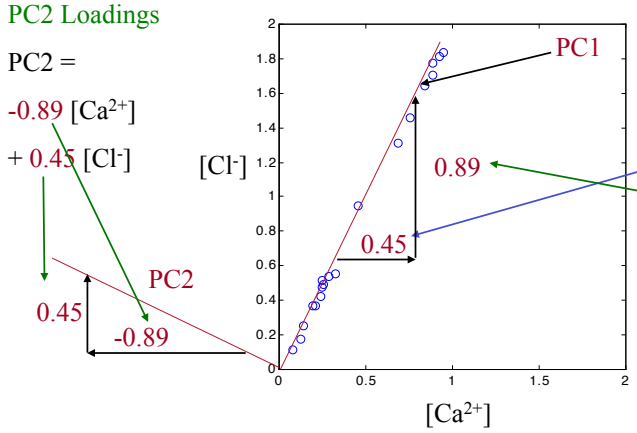
Linear Combination of Original Axes

PC2 Loadings

PC2 =

$-0.89\ [Ca^{2+}]$

$+ 0.45\ [Cl^-]$



PC1 =

$0.45\ [Ca^{2+}]$

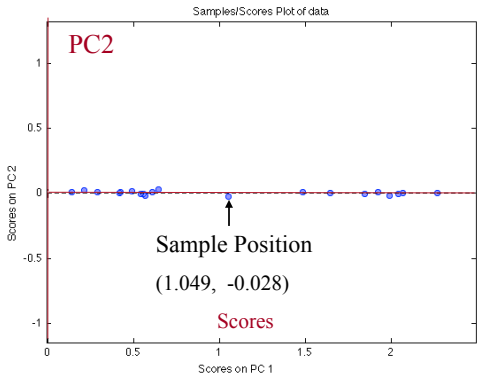
$+ 0.89\ [Cl^-]$

Loadings

36

EIGENVECTOR RESEARCH INCORPORATED

The Positions of the Data in the PC Coordinate System (Scores)



Same as last plot except rotated to make PCs new axes

(PCs are a new coordinate system)

PC1

EIGENVECTOR RESEARCH INCORPORATED

Summary of Ca^{2+}/Cl^- Example

$$PC1 = 0.45\,[Ca^{2+}] + 0.89\,[Cl^-]$$

- Loadings for PC1 in a ratio of 1:2
  - One part Ca^{2+} and two parts Cl^-
- PC1 related to $[CaCl_2]$
  - ~ 99.97% variance is systematic
- PC2 related to random measurement error
  - ~ 0.03% noise

EIGENVECTOR RESEARCH INCORPORATED

Important Mathematical Term: Chemical Rank of the Matrix:

The number of independent, underlying, meaningful sets of information in a data set.*

- Calcium ion, chloride ion example
  - Started with two variables: $[Ca^{+2}]$ and $[Cl^{-}]$
  - Ended with one meaningful variable: $CaCl_2$
  - Chemical Rank = 1

*This is really the definition of "pseudo-rank." The mathematical rank is the number of linearly independent rows or columns.

EIGENVECTOR RESEARCH INCORPORATED

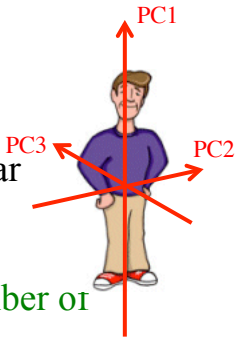
Summary of What We Have Done and Learned

- Variance = Information
- Many variables contain information, but much of the information may be redundant (correlated) or irrelevant.
- Use linear combinations of the original variables to create new variables (Principal Components, PCs) that combines redundant information → PCA
- The Rank of a Matrix is the number of Principal Components that describe other than random noise.

EIGENVECTOR RESEARCH INCORPORATED

How Does PCA Find the PCs?

- The 1st principal component (PC) passes through the origin and the maximum variance of the data.
- The 2nd PC is orthogonal (perpendicular or independent) to PC 1 and passes through the second most variance.
- The process is continued until the number of new PCs = number of old variables.
  - PCs ≤ min(number of samples, number of variables) = mathematical rank of the data



EIGENVECTOR RESEARCH INCORPORATED

What Does PCA Give Me?

- Most of the variance (information) is concentrated in the first few PCs.
  - Some may be relevant to the problem of interest
- Small random noise is sifted into the later PCs
  - and may be thrown away - data filtering
  - or used in a residuals analysis
- Important Assumption:
  - The signal/noise is > 1
  - *i.e.,* most of the variance is from sources other than random noise

EIGENVECTOR RESEARCH INCORPORATED

What Does PCA Give Me?

- Loadings: Compositions of the new PC axes in terms of the old variables. May be able to interpret the loadings in chemical terms.
  - Loadings ⟷ Variables
- Scores: The position of the samples in the new PC coordinate system. The closer samples are to each other in the first few PC space, the more they are alike.
  - Scores ⟷ Samples
- Eigenvalues - The variance stored in each of the Principal Components
  - Eigenvalues can then be used to calculate the % of the information stored in each PC.

43

EIGENVECTOR RESEARCH INCORPORATED

Say 100 Times Before You Go to Sleep Tonight!!

Loadings ⟷ Variables

Scores ⟷ Samples

Eigenvalues ⟷ Variance

44

EIGENVECTOR RESEARCH INCORPORATED

Arch Example: a Little More Complicated Problem

Track trading and migration patterns of prehistoric tribes. Samples of two unknown obsidian artifacts were compared to samples obtained from three obsidian quarries using XRF.

Quarry	Number of Samples	Row Index
K	10	# 1-10
SH	23	#11-33
AN	21	#34-54
Unknowns	2	#55 & 56
Total	56	

BR Kowalski, TF Schatzki, FH Stross. "Classification of archaeological artifacts by applying pattern recognition to trace element data." *Anal. Chem.*, **44**(13); 2176-2180 (1972).

EIGENVECTOR RESEARCH INCORPORATED

20 Obsidian Samples Were Analyzed for 10 Metals*

	Fe	Ti	Ba	Ca	K	Mn	Rb	Sr	Y	Zr
1	1173	417	54	961	441	47	135	55	60	145
2	1164	404	56	916	446	42	120	58	45	148
3	1030	373	59	920	487	38	128	53	58	138
4	1077	373	55	888	455	38	97	51	54	145
5	1080	403	53	919	442	41	133	60	45	155
•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•
56	880	156	36	279	388	37	103	15	53	143

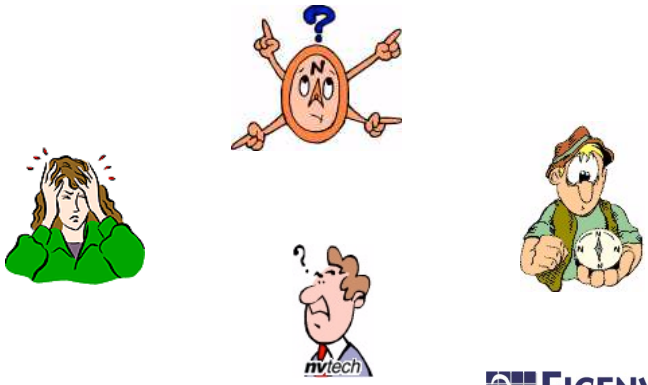
56 x 10 Matrix

*X-Ray Fluorescence (ppm)

D.F. Stevenson, et. al., *Archaeometry*, **13**, 17 (1971).

EIGENVECTOR RESEARCH INCORPORATED

Plot Each Sample as a Point in 10-Dimensional Space

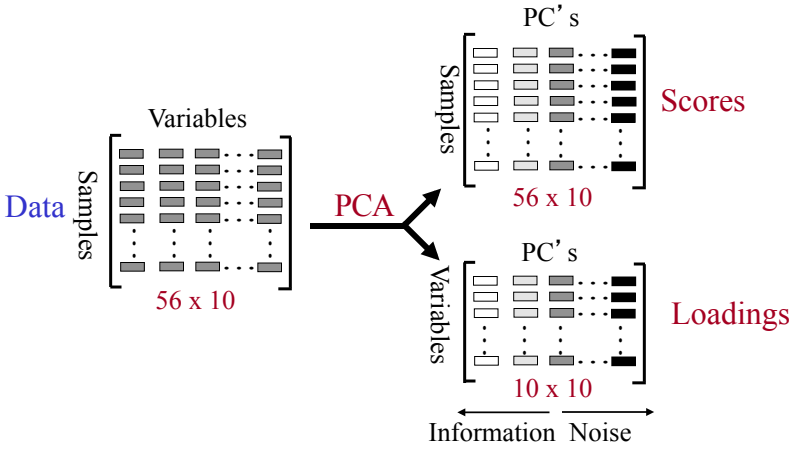


47

EIGENVECTOR RESEARCH INCORPORATED

Let's Do It the Right Way

PCA



48

EIGENVECTOR RESEARCH INCORPORATED

When Washing Clothes
Sometimes Detergent is Not Enough

You may need Pretreatments

And You Need a Different Pretreatment for Different Stains

49

EIGENVECTOR RESEARCH INCORPORATED

Before PCA do:
Data Preprocessing

- None
- Mean centering
- Autoscaling

50

EIGENVECTOR RESEARCH INCORPORATED

PCA of GLASS - No Preprocessing

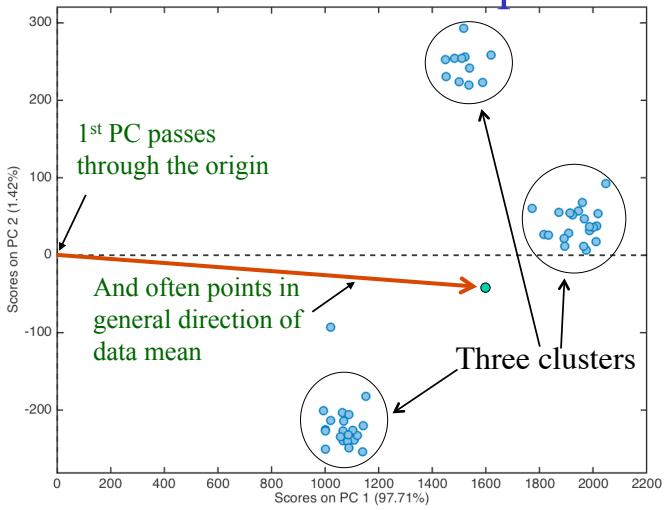
```
Principal Components Analysis Model
X-block: arch  56 by 10
Included: [ 1-56 ]  [ 1-10 ]
Preprocessing: None
Num. PCs: 5
```

Percent Variance Captured by PCA Model

Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	2.36e+06	97.71	97.71
2	3.44e+04	1.42	99.13
3	1.93e+04	0.80	99.93
4	1.24e+03	0.05	99.98
5	2.08e+02	0.01	99.99

EIGENVECTOR RESEARCH INCORPORATED

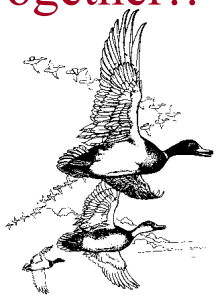
PCA Scores Plot - No Preprocessing



EIGENVECTOR RESEARCH INCORPORATED

If we know and label the source of the known samples (Learning Set), we can identify the clusters and the memberships of the unknowns.

Birds of a Feather Flock Together!!

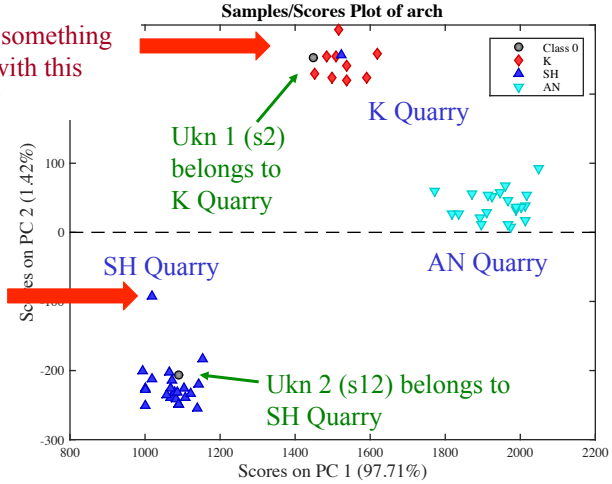


53

EIGENVECTOR RESEARCH INCORPORATED

PCA Scores Plot - No Preprocessing

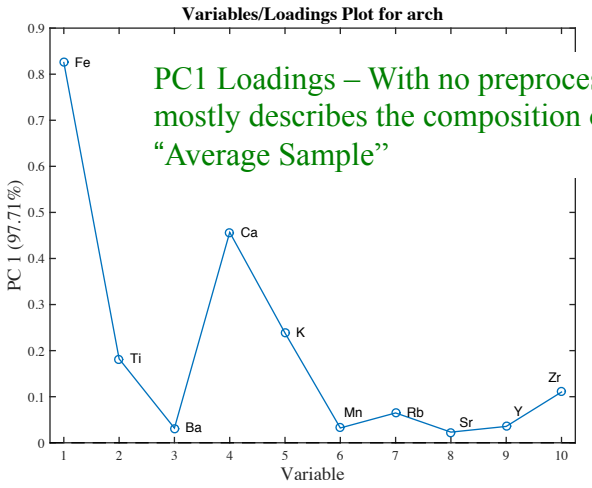
Is there something wrong with this picture?



54

EIGENVECTOR RESEARCH INCORPORATED

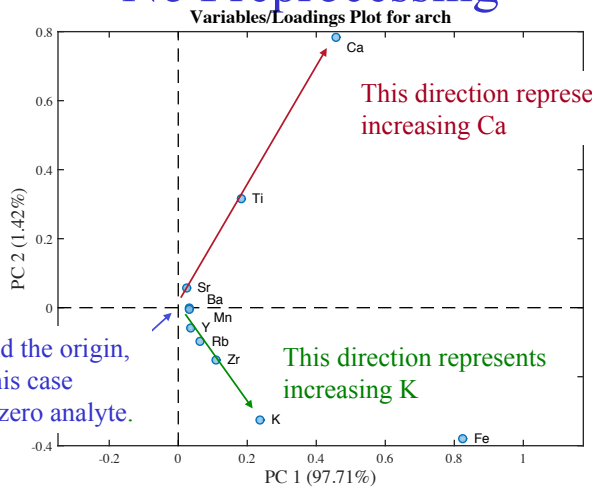
Loadings Plot



PC1 Loadings – With no preprocessing it mostly describes the composition of the "Average Sample"

EIGENVECTOR RESEARCH INCORPORATED

PCA Loading Plot PC 2 vs PC 1 - No Preprocessing



This direction represents increasing Ca

Always find the origin, which in this case represents zero analyte.

This direction represents increasing K

EIGENVECTOR RESEARCH INCORPORATED

Loadings Plots are like a Signpost to Locations in the familiar world of measured variables

57

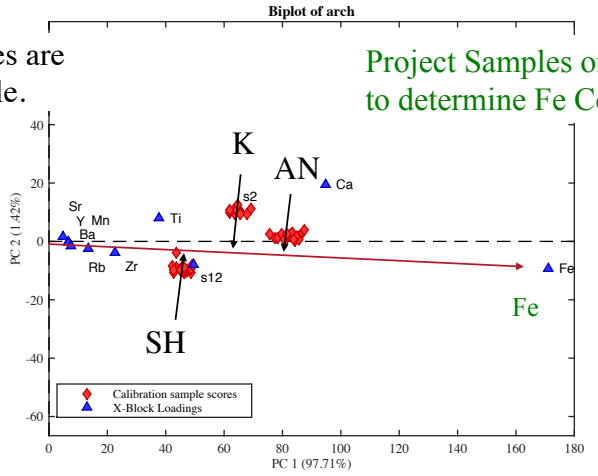
EIGENVECTOR RESEARCH INCORPORATED

Easier to Understand by Superimposing Scores and Loadings - Biplot

Note the axes are equal in scale.

Project Samples onto Fe line to determine Fe Content.

PC 1+2 ~99.1%



58

EIGENVECTOR RESEARCH INCORPORATED

Interpret this Biplot:

- According to the first two PCs:
  - the AN Quarry has the highest Fe content,
  - the SH Quarry has the lowest Fe content, and
  - the K Quarry has ~average Fe content.
- Difference in Fe content is one source of discrimination between the quarries.

What we have learned

- The first two PCs can discriminate between the three quarries.
- PC 1 contains 99% variance, but mostly describes what the quarries have in common.
- But we really want to understand the difference between the quarries…

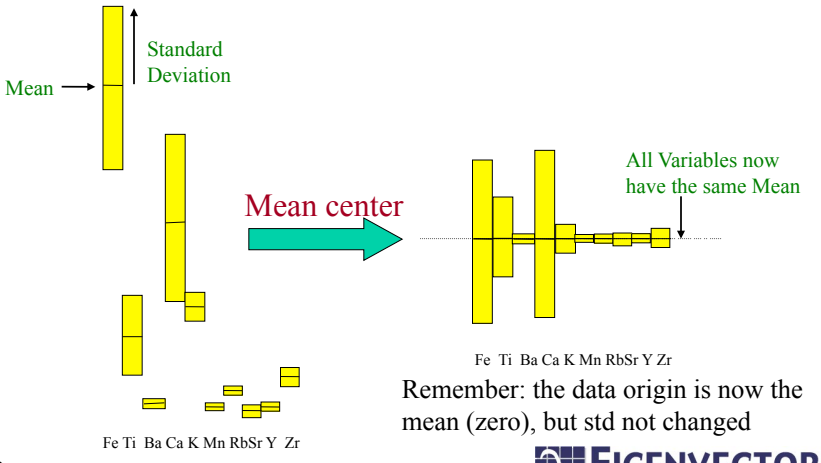
Data Preprocessing

- None
- Mean centering
- Autoscaling

61

EIGENVECTOR RESEARCH INCORPORATED

Repeat the PCA with Mean Centered Data

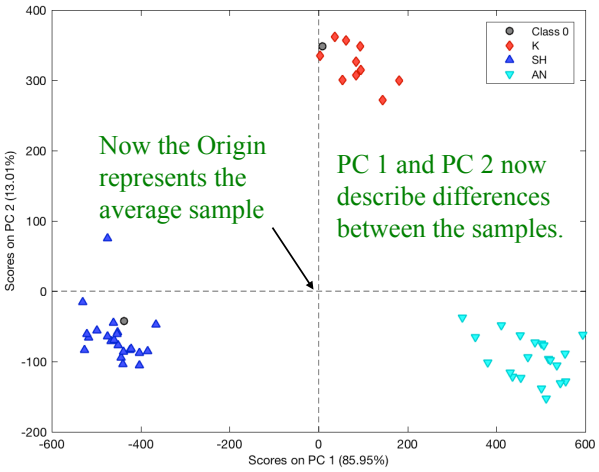


Remember: the data origin is now the mean (zero), but std not changed

62

EIGENVECTOR RESEARCH INCORPORATED

Scores Plot - Mean Centering

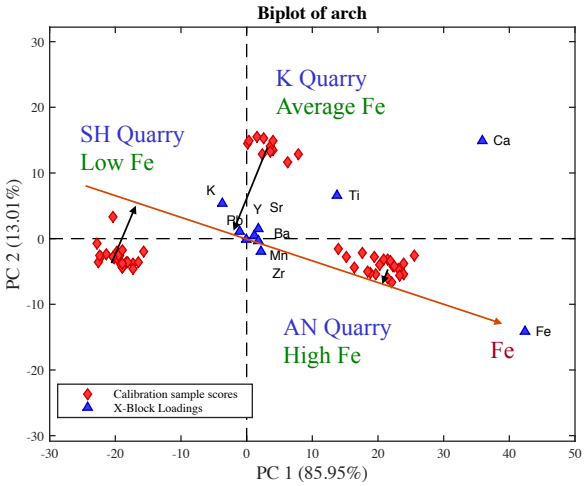


PC 2 13% of Variance

PC 1 86% of Variance

EIGENVECTOR RESEARCH INCORPORATED

Biplot - Mean Centering



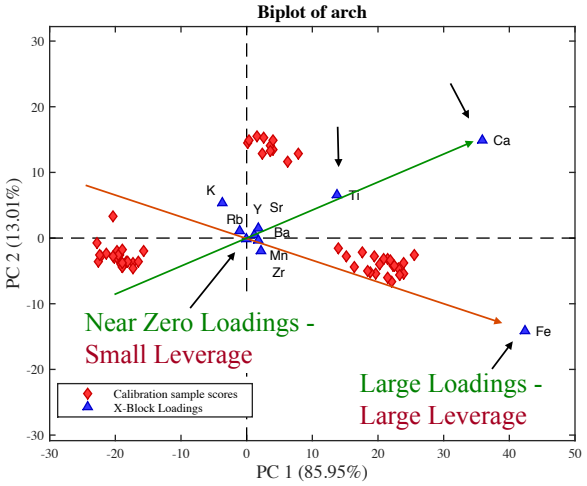
Note axes have equal scale

PC 1+2 ~98.97%

Fe is important in discriminating between the Quarries in PC 1 - PC 2 space.

EIGENVECTOR RESEARCH INCORPORATED

Biplot - Mean Centering



Fe, Ca and Ti are important in discriminating between the Quarries in PC 1 - PC 2 space.

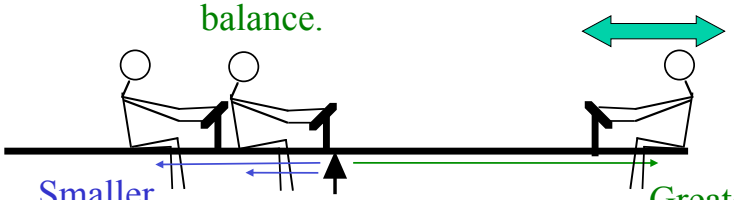
Large Leverage

65

EIGENVECTOR RESEARCH INCORPORATED

Leverage: Distance from Origin

Greater Leverage because the slightest change in position will greatly upset the balance.



Smaller Leverages

Origin

Greater Leverage

Leverage is measured by a quantity T^2.

66

EIGENVECTOR RESEARCH INCORPORATED

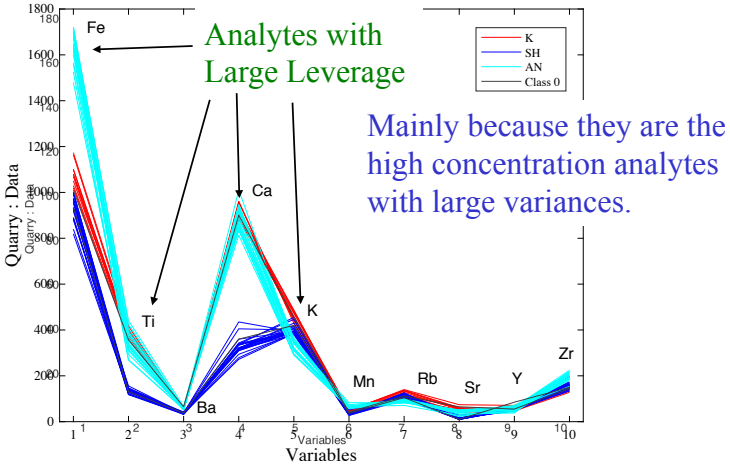
Variables with Small Leverages:

- Have little to say about the differences between the three quarries.
- And/or have relatively small signals (*i.e.*, low concentrations) compared to signals from other variables.
- Both of these situations result in small variances - little information

67

EIGENVECTOR RESEARCH INCORPORATED

Plot the Data: Analyte Profile



68

EIGENVECTOR RESEARCH INCORPORATED

Do all four participants in this decision have the same Leverage?

Much More Fair!

69

EIGENVECTOR RESEARCH INCORPORATED

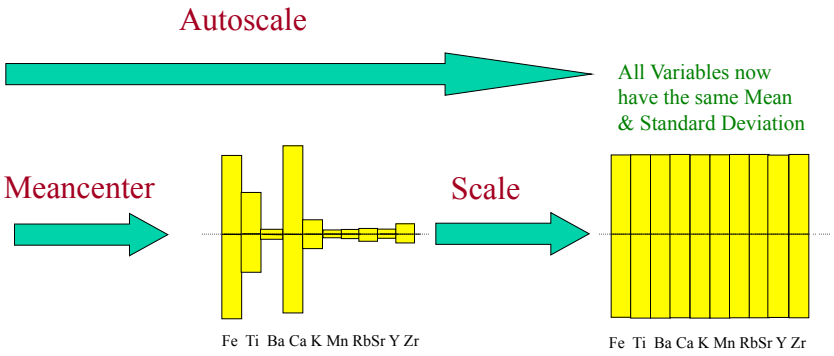
Data Preprocessing

- None
- Mean centering
- Autoscaling

70

EIGENVECTOR RESEARCH INCORPORATED

Autoscaling

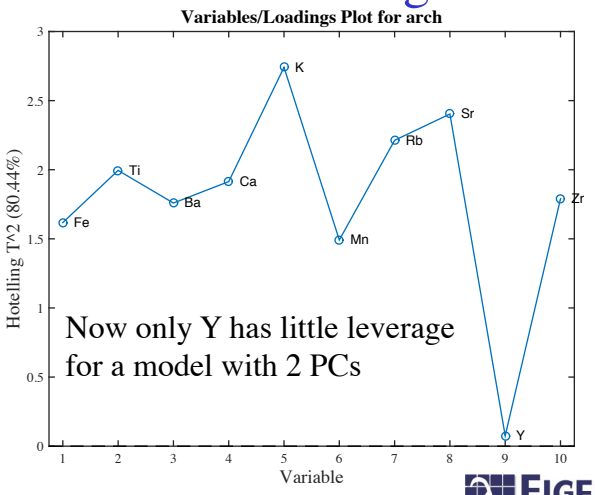


Remember: autoscaling includes mean centering

71

EIGENVECTOR RESEARCH INCORPORATED

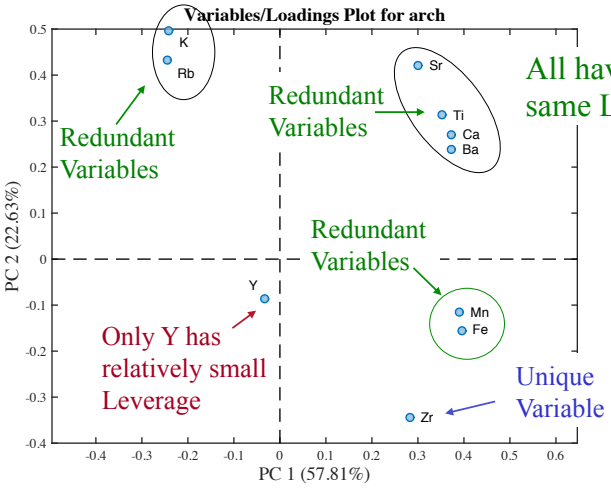
Leverages of Variable after Autoscaling



72

EIGENVECTOR RESEARCH INCORPORATED

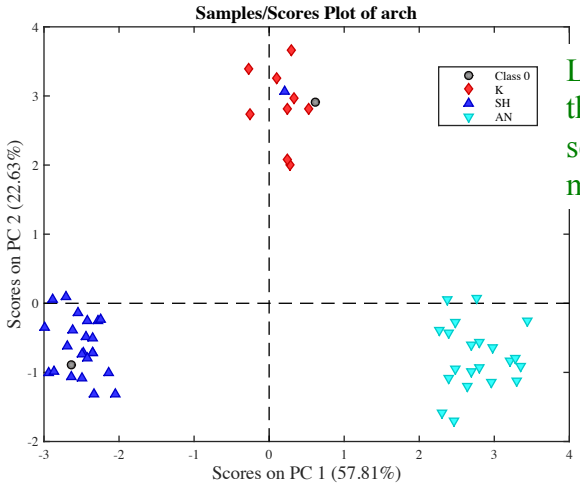
Loadings Plot - Autoscaling



73

EIGENVECTOR RESEARCH INCORPORATED

PCA Scores Plot -Autoscaling

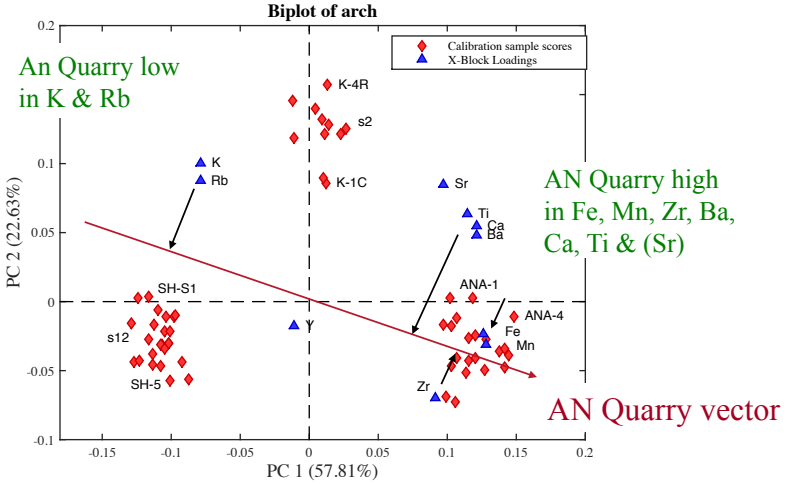


Looks much the same as the scores plot for mean centering

74

EIGENVECTOR RESEARCH INCORPORATED

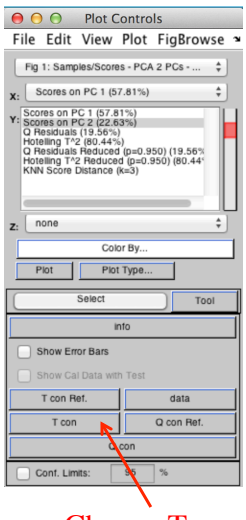
Biplot - Autoscaling

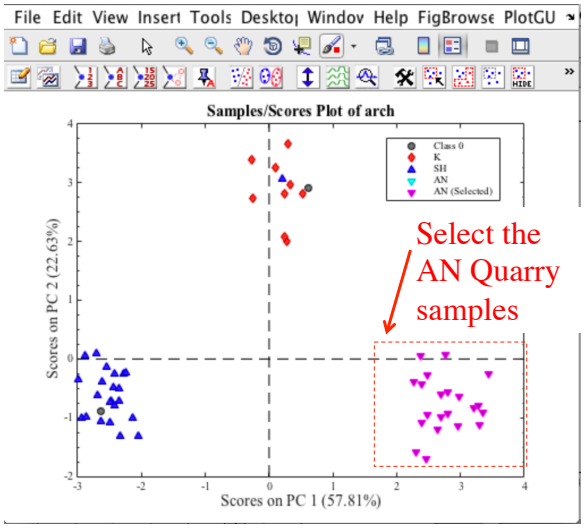


75

EIGENVECTOR RESEARCH INCORPORATED

T- Contribution of Variables to Samples

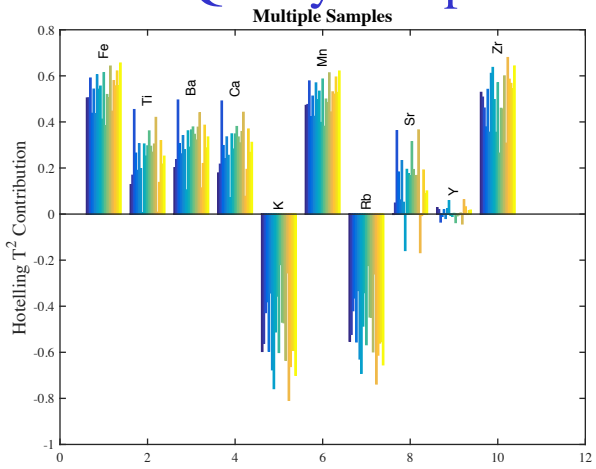




76

EIGENVECTOR RESEARCH INCORPORATED

Variable Leverage of AN Quarry Samples



77

EIGENVECTOR RESEARCH INCORPORATED

What Can I See From Biplot or T-Con?

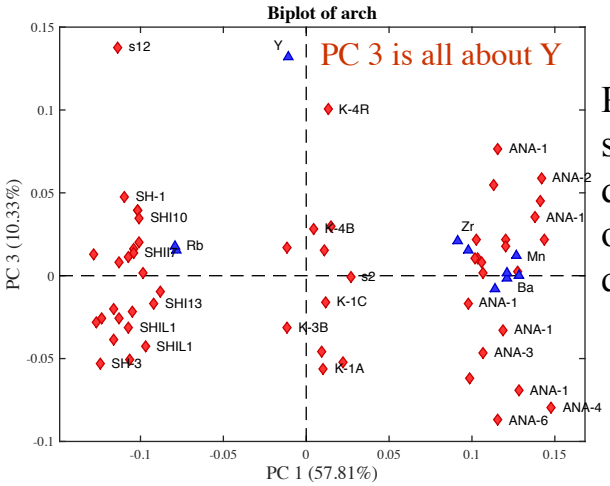
(In 2 PC - Autoscaled space)

- AN Quarry:
  - High in Fe, Mn & Zr
  - Moderately high in Ba, Ca, Ti & (Sr)
  - Low in K & Rb
- SH Quarry:
  - Low in Fe, Ti, Ba, Ca, Mn & Sr
- K Quarry:
  - High in K, Rb & Sr
  - Moderately high in Ti, Ba & Ca
  - Low in Zr

78

EIGENVECTOR RESEARCH INCORPORATED

Also Check Higher PCs



PC 3 Scores show no discrimination of samples by quarry

79

EIGENVECTOR RESEARCH INCORPORATED

When Should We Mean Center or Autoscale?

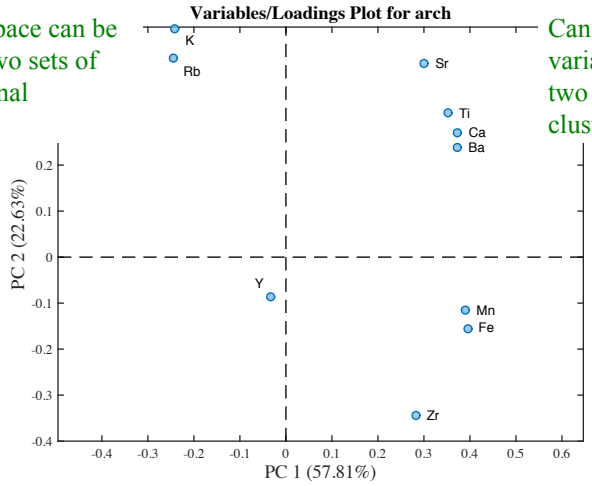
- In exploratory analysis, nearly always mean center
  - Remember autoscaling includes mean centering.
- Autoscale when
  - variables are of different units *e.g.,* cm, g, C, atm, etc.
  - you are confident that low-signal variables still have good signal-to-noise ratios
  - not often in spectra - usually all same units *e.g.* abs
- Autoscaling weights each variable equally
  - Don't want to scale up noise.
- When in doubt, try it and see what happens

80

EIGENVECTOR RESEARCH INCORPORATED

Examine Loadings Plot Again

Much of this space can be described by two sets of nearly orthogonal variables.

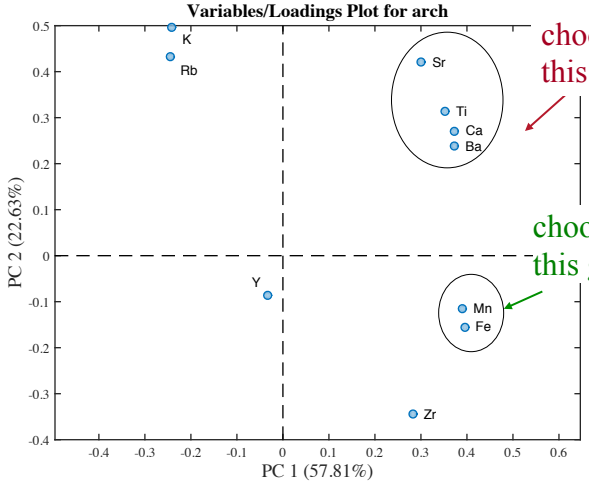


Can choose one variable from each of two ~ orthogonal clusters.

81

EIGENVECTOR RESEARCH INCORPORATED

Examine Loadings Plot Again



choose Ca from this group

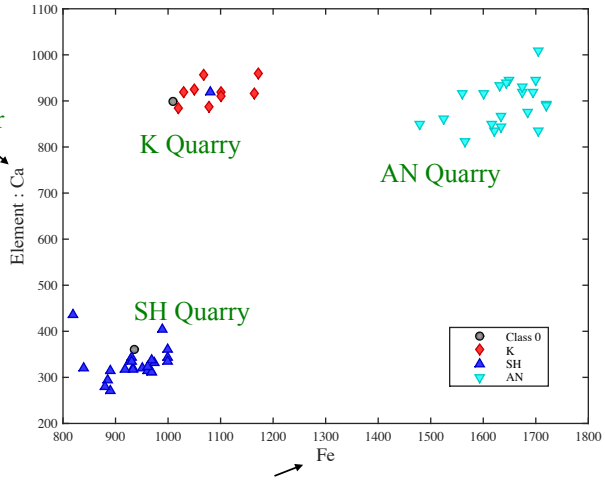
choose Fe from this group

82

EIGENVECTOR RESEARCH INCORPORATED

Plot [Ca] vs. [Fe] - Variable Selection

Could also use Sr, Ti or Ba here



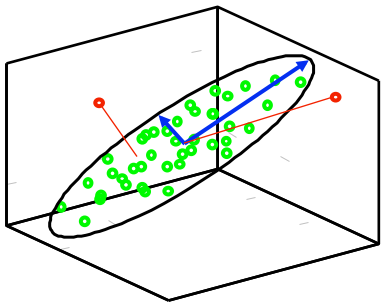
Could also use Zr, Mn, Rb or K

83

EIGENVECTOR RESEARCH INCORPORATED

What Have We Accomplished?

- Concentrated discriminating power of 10 variables into 2 PCs containing 80% of the information using autoscaling
- Have determined characteristics of each quarry that differentiates it from other quarries.
- If lab analyses were difficult and/or expensive, found that we can differentiate quarries with just 2 variables.



84

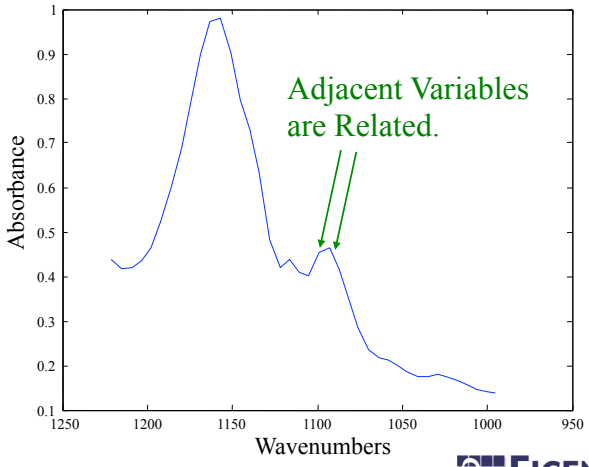
EIGENVECTOR RESEARCH INCORPORATED

Yet a Little More Complicated Problem

- We wish to use FT-IR spectra and pattern recognition to distinguish authentic olive oil from counterfeit or adulterated olive oil.
- This time we shall see some special properties associated with Spectral Data.
- We shall also learn a new pattern recognition technique called SIMCA.

EIGENVECTOR RESEARCH INCORPORATED

What is Special about Spectral Data?



EIGENVECTOR RESEARCH INCORPORATED

- PCA treats each variable as unrelated and unordered until it discovers relationships between variables.
- We could rearrange the order of the variables and PCA would not care.
- As chemists we know that absorbance in a spectrum must contain smoothly varying values.
- This will lead to special pretreatment techniques not available to other types of data.

EIGENVECTOR RESEARCH INCORPORATED

Olive Oil Samples
Learning Set:

Sample	Number of Samples	Sample Indices
Corn Oil	9	# 1-9
Olive Oil	15	# 10-24
Safflower Oil	8	# 25-32
Corn Margarine	4	# 33-36

Took FT-IR spectra (3600 - 600 cm^{-1}) of these oils using a fixed pathlength NaCl cell.

EIGENVECTOR RESEARCH INCORPORATED

Test Set:

Sample	No. Samples	Indices
Corn Oil*	9	# 1-9
Olive Oil*	15	# 10-24
Safflower Oil*	8	# 25-32
Corn Margarine*	4	# 33-36
Corn Oil in Olive Oil	5	# 37-41
5, 10, 20, 30 & 40%		
Almond Oil	1	(#42)
Peanut Oil	1	(#43)
Sesame Oil	1	(#44)

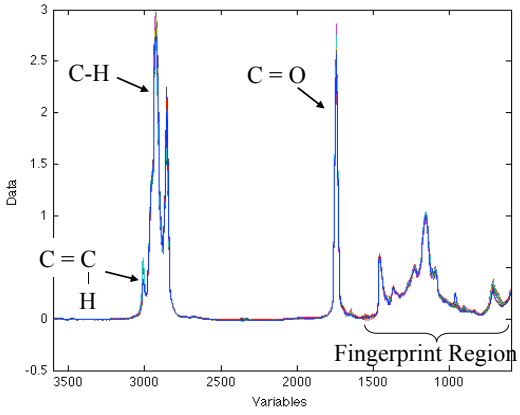
* New Samples

89

EIGENVECTOR RESEARCH INCORPORATED

FTIR Spectra of 36 Sample Learning Set

Notice how spectra look alike.



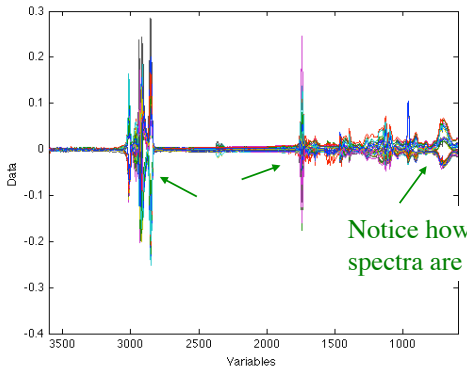
90

D.B. Dahlberg, et. al, *Applied Spectroscopy*, **51**, 1118 (1997).

EIGENVECTOR RESEARCH INCORPORATED

Do PCA on the Learning Set

- Examine full wavelength range
- Mean-center

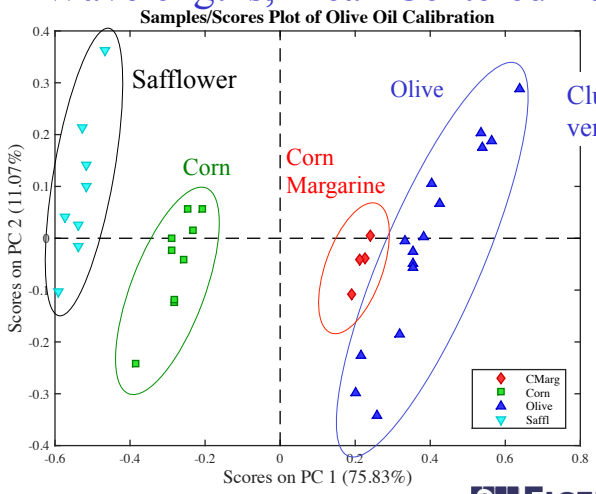


Notice how differences between spectra are emphasized

EIGENVECTOR RESEARCH INCORPORATED

Scores Plot

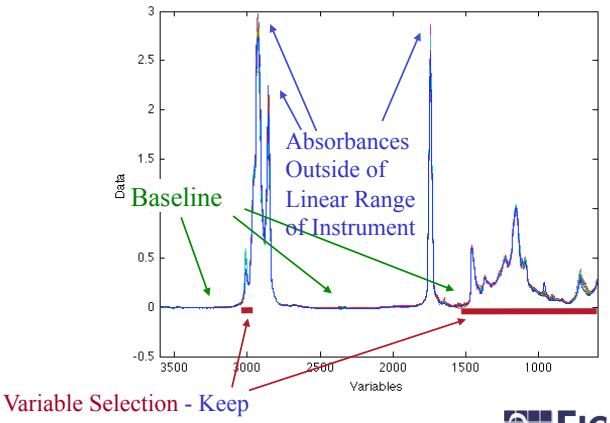
All Wavelengths, Mean Centered Data



Clusters are not very tight

EIGENVECTOR RESEARCH INCORPORATED

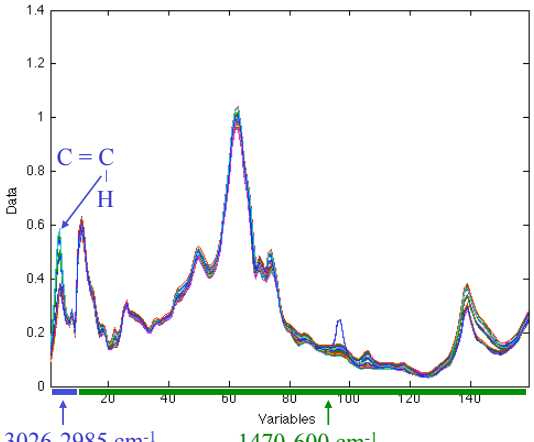
How Can the Model be Improved?



93

EIGENVECTOR RESEARCH INCORPORATED

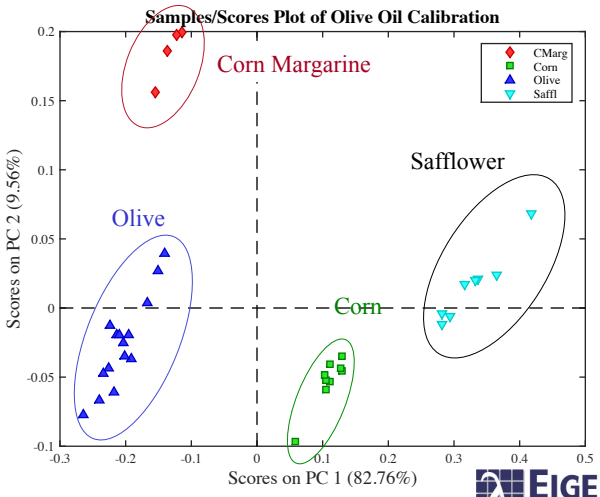
FTIR Spectra of Learning Set Selected Variables



94

EIGENVECTOR RESEARCH INCORPORATED

Scores Plot
Selected Wavelengths, Mean Centered Data

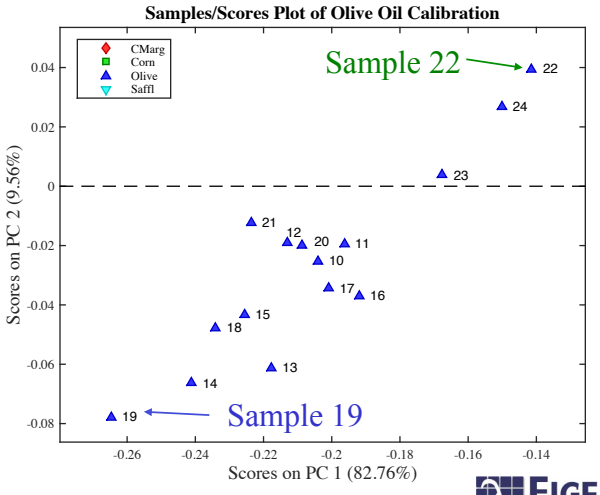


Better, but still too much spread

95

EIGENVECTOR RESEARCH INCORPORATED

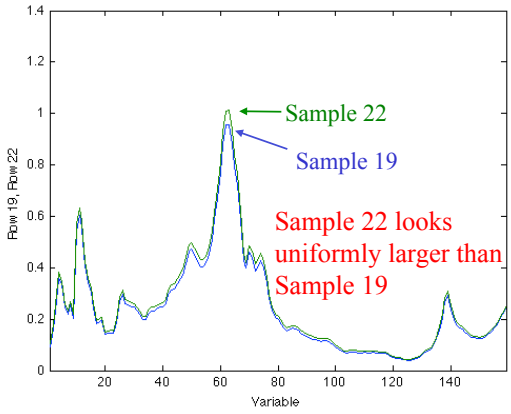
Scores Plot
Zoom In on Olive Samples



96

EIGENVECTOR RESEARCH INCORPORATED

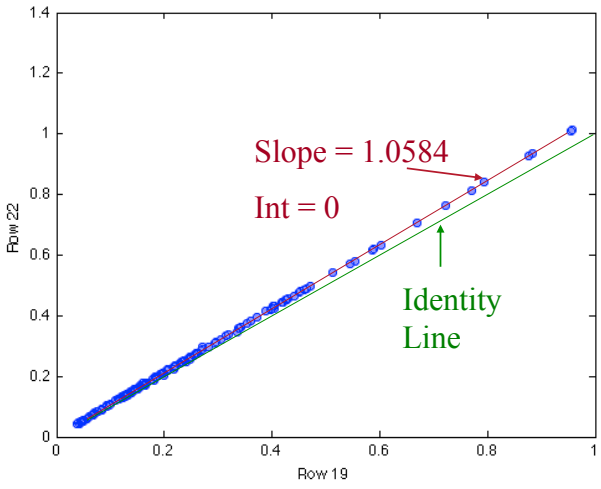
Spectra (Selected Wavelengths)
Samples 19 & 22



97

EIGENVECTOR RESEARCH INCORPORATED

Plot Sample 22 vs. Sample 19



98

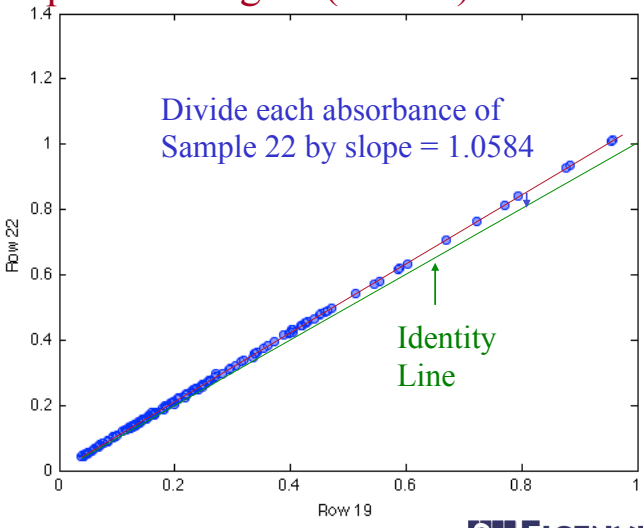
EIGENVECTOR RESEARCH INCORPORATED

Multiplicative Effect
Two Spectra are Identical except one is a Multiple of the Other

- Changing sample pathlength, *e.g.* changing light scattering with particle size.
- Changing sample density, *e.g.* changing temperature of sample.
- Changing gain of the instrument.

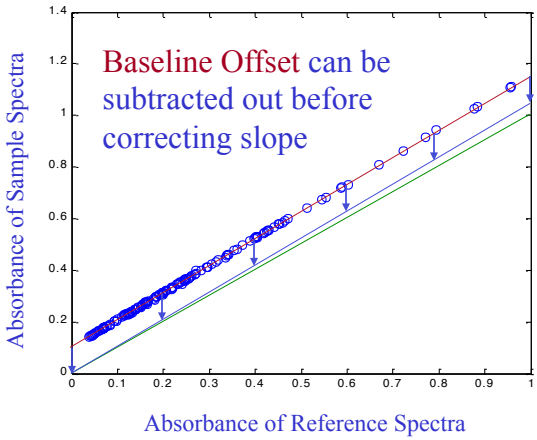
EIGENVECTOR RESEARCH INCORPORATED

MSC
Multiplicative Signal (Scatter) Correction



EIGENVECTOR RESEARCH INCORPORATED

If there is also an Offset



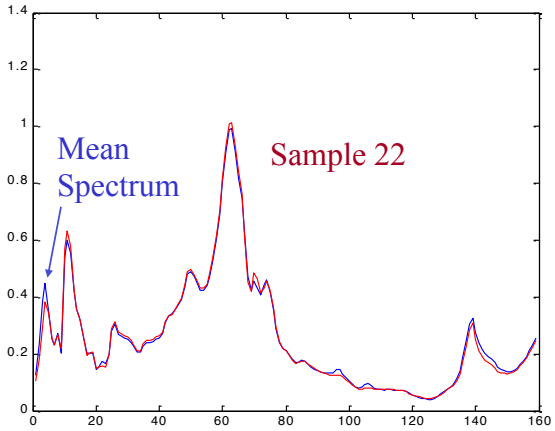
EIGENVECTOR RESEARCH INCORPORATED

What to use as a Reference Spectrum?

- Anything we want that looks like the spectra in the Learning Set.
- Usually choose the Mean Spectrum of the Learning Set.
  - The same spectrum subtracted when doing mean centering of all samples, learning set, test set and real samples.

EIGENVECTOR RESEARCH INCORPORATED

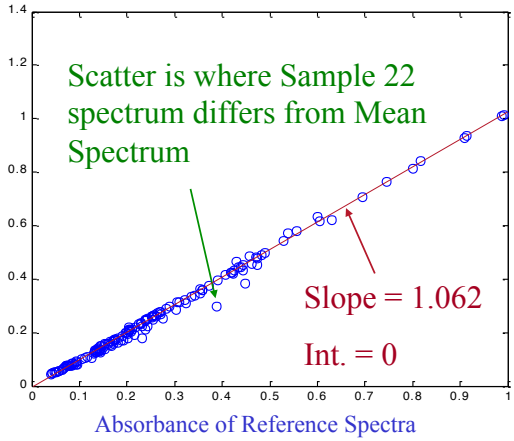
Mean Spectrum and Sample 22



103

EIGENVECTOR RESEARCH INCORPORATED

Sample 22 vs. Mean Spectrum



104

EIGENVECTOR RESEARCH INCORPORATED

Applying MSC

- Software will do MSC on each sample of the Learning Set
- You need to specify the Learning Set and the Reference Spectrum
  - usually mean spectrum of Learning Set

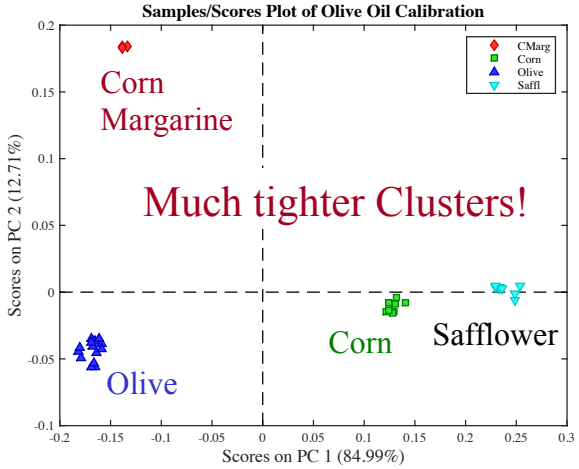
Geladi P, MacDougall D, Martens H., *Appl. Spectrosc.*, **39**(3), 491-500 (1985)

105

EIGENVECTOR RESEARCH INCORPORATED

Scores Plot

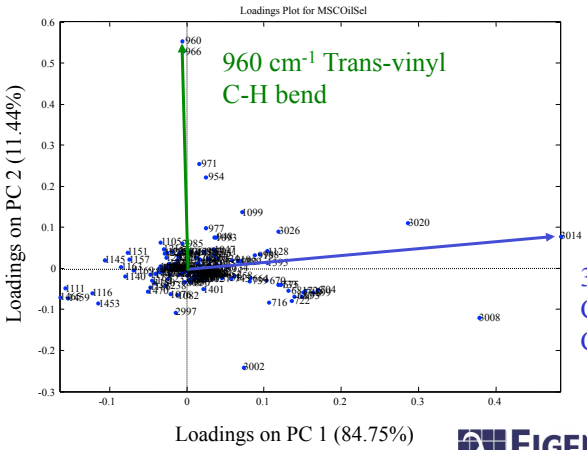
Selected Wavelengths, MSC & Mean Centered



106

EIGENVECTOR RESEARCH INCORPORATED

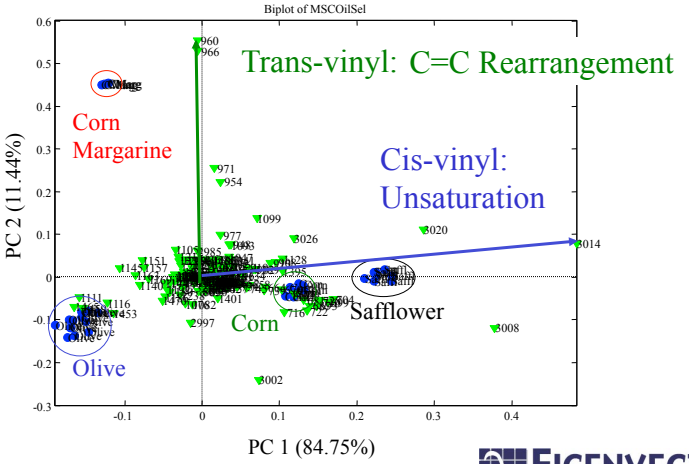
Loadings Plot
Selected Wavelengths, MSC & Mean Centered



107

EIGENVECTOR RESEARCH INCORPORATED

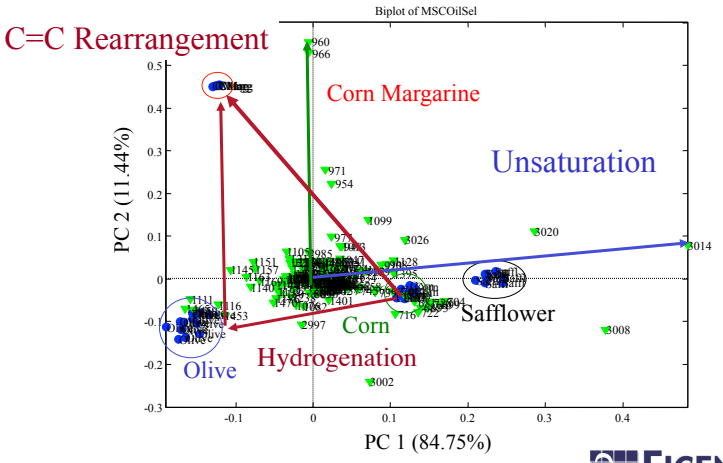
Biplot
Selected Wavelengths, MSC & Mean Centered



108

EIGENVECTOR RESEARCH INCORPORATED

Biplot
Selected Wavelengths, MSC & Mean Centered



109

EIGENVECTOR RESEARCH INCORPORATED

Pattern Recognition

- Which samples are most like/unlike other Samples (Scores)
  - Outlier detection
  - Sample(s) most like target sample
  - Cluster analysis
- Which variables are most like/unlike other Variable (Loadings)
  - Variable selection
  - Detect underlying phenomena

110

EIGENVECTOR RESEARCH INCORPORATED

Pattern Recognition Can Relate Samples and Variables (Biplot)

- Determine which variables are responsible for sample similarities and differences.
  - Yields knowledge about a system
  - Leads to improvements in the system process, product, experimental improvements

111

EIGENVECTOR RESEARCH INCORPORATED

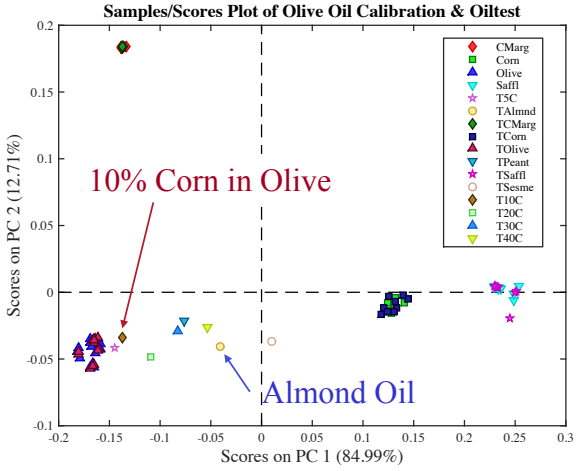
Can PCA be used to Detect Substitute or Adulterated Olive Oils?

- Bring in the Test Set
  - New samples of Olive, Corn, Corn Marg. & Safflower
  - Other Oils
  - 5, 10, 20, 30 & 40% Corn Oil in Olive Oil
- MSC
- Mean center
- PCA

112

EIGENVECTOR RESEARCH INCORPORATED

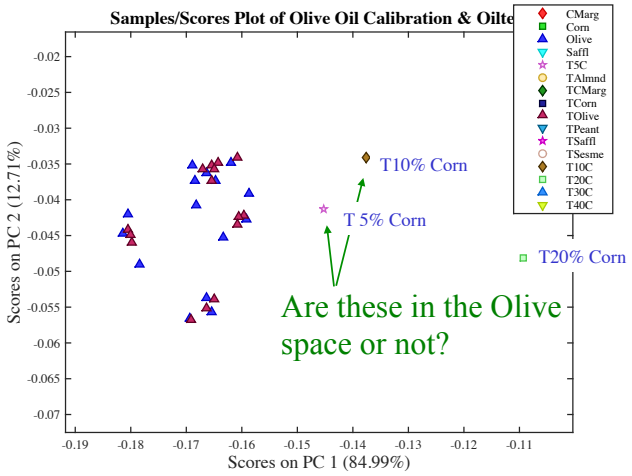
Score Plot Learning Set plus Unknowns "T"



113

EIGENVECTOR RESEARCH INCORPORATED

Closer View of Olive Oils



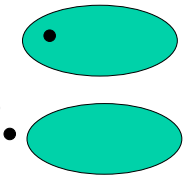
114

EIGENVECTOR RESEARCH INCORPORATED

Need a method to put statistical Boundaries around the Olive Oil

If a sample falls within the boundaries, it is Olive Oil.

If a sample falls outside the boundaries, it is not Olive Oil.



One Technique is:

Soft Independent Method of Class Analogy

SIMCA

115

EIGENVECTOR RESEARCH INCORPORATED

Outline (Part 1)

- Introduction
- Pattern Recognition Motivation
- Principal Components Analysis
- SIMCA
  - Soft Independent Method Class Analogy
- Summary

116

EIGENVECTOR RESEARCH INCORPORATED

How SIMCA works

- Perform PCA on a Learning Set representing ONE class, *e.g.* Olive Oil
- Verify that you have a proper Learning Set.
  - samples are representative of all olive oils to be classed

How?

- Choose the number of PCs that are sufficient to describe this set.
  - This is the Model of that class.
- Set statistical Confidence Limits, *i.e.* boundaries for the Model.
- Determine if an unknown lies within or outside these limits.
- Repeat this process for Each class.

117

EIGENVECTOR RESEARCH INCORPORATED

Determine the Number of PCs That Are Sufficient to Describe Olive Oil

Perhaps You Know Your Analytical Technique is Good to ± 10%

Where does the Model end and the noise begin?

What is the Rank of the data?

	Eigenvalue of Cov(X)	% Variance This PC	% Variance Cumulative
1	2.72e-04	40.43	40.43
2	2.07e-04	30.80	71.23
3	7.08e-05	10.51	81.74
4	6.58e-05	9.77	91.51
5	2.71e-05	4.02	95.53
6	1.02e-05	1.51	97.04
7	9.45e-06	1.40	98.45
8	3.30e-06	0.49	98.94
9	2.67e-06	0.40	99.33
10	1.64e-06	0.24	99.58
11	9.75e-07	0.14	99.72
12	7.73e-07	0.11	99.84
13	5.69e-07	0.08	99.92
14	5.31e-07	0.08	100.00

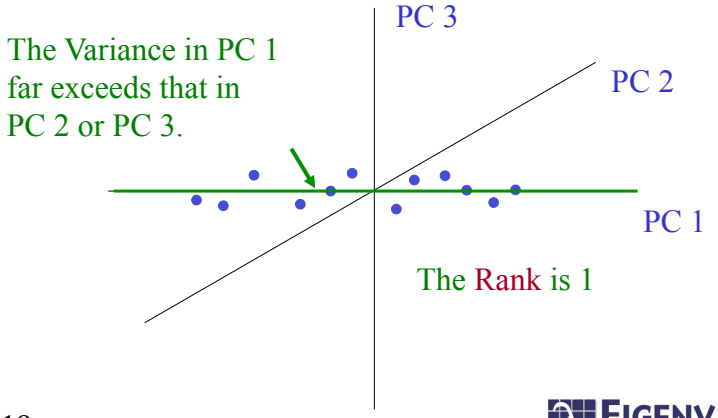
Terminate Model with 3 PCs

Beginning to Model 10% Noise

118

EIGENVECTOR RESEARCH INCORPORATED

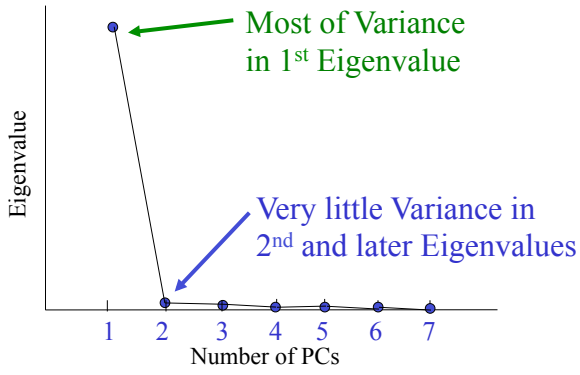
Another Approach is to Examine Distribution of Variance among the PCs



119

EIGENVECTOR RESEARCH INCORPORATED

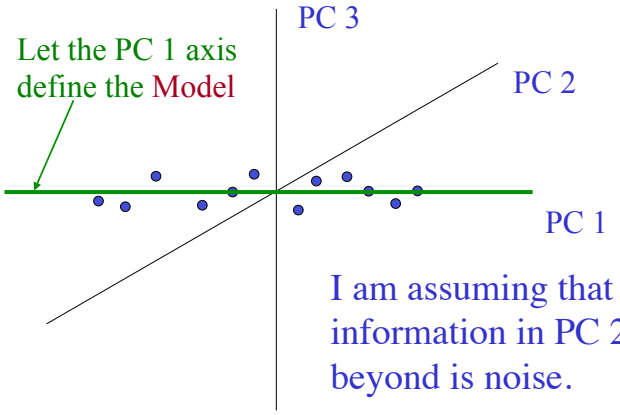
Plot Eigenvalues



120

EIGENVECTOR RESEARCH INCORPORATED

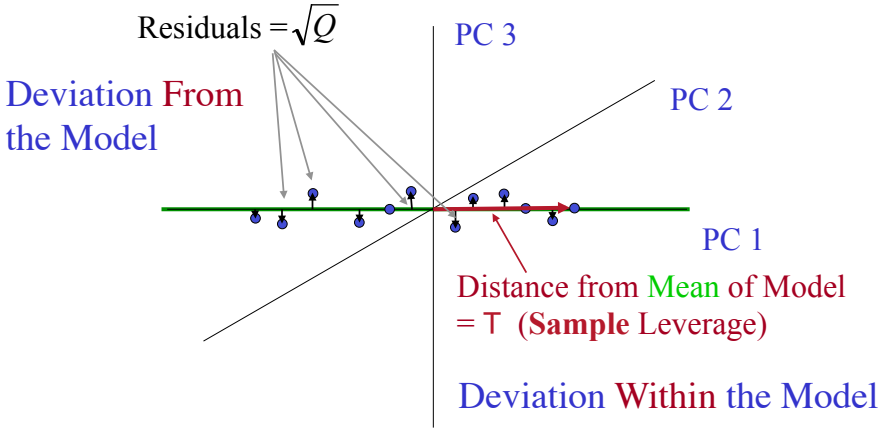
Create One PC Model with Boundaries



121

EIGENVECTOR RESEARCH INCORPORATED

Individual Points Deviate from the Group in Two Ways



122

EIGENVECTOR RESEARCH INCORPORATED

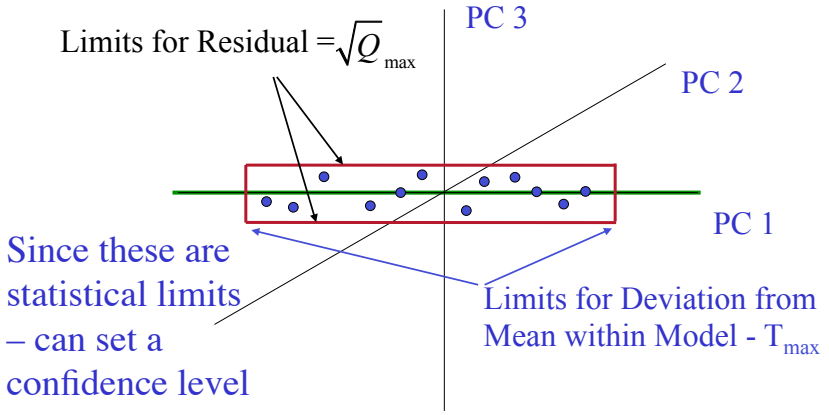
Methods of Measuring Distance

- Euclidean Distance: Shortest distance between two samples. Here we're interested in the distance from the sample to the mean.
- Mahalanobis Distance: Distance scaled along each axis by dividing by the Standard Deviation of that axis corrected for the covariance between the axes. (this is also how T^2 is calculated).
- Mahalanobis Distance has several statistical advantages over Euclidean Distance.

123

EIGENVECTOR RESEARCH INCORPORATED

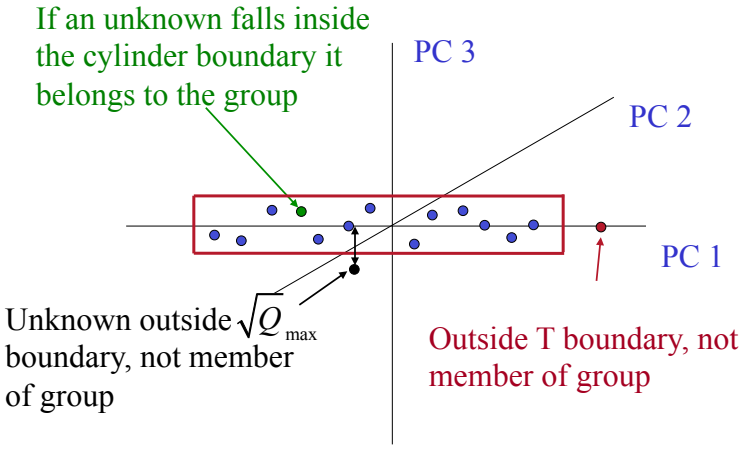
Set Class Boundaries Based on Deviation of Learning Set from Idealized Model



124

EIGENVECTOR RESEARCH INCORPORATED

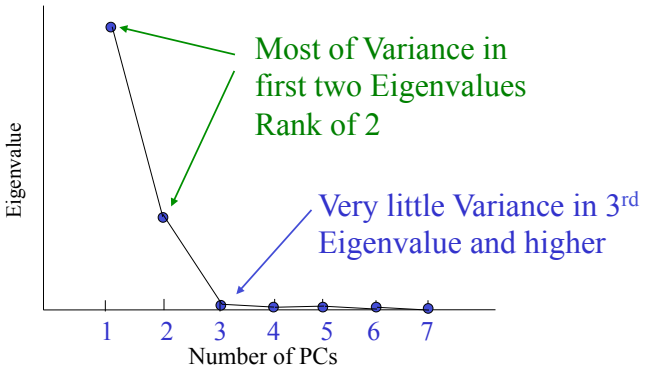
Place Unknown into PC Space



125

EIGENVECTOR RESEARCH INCORPORATED

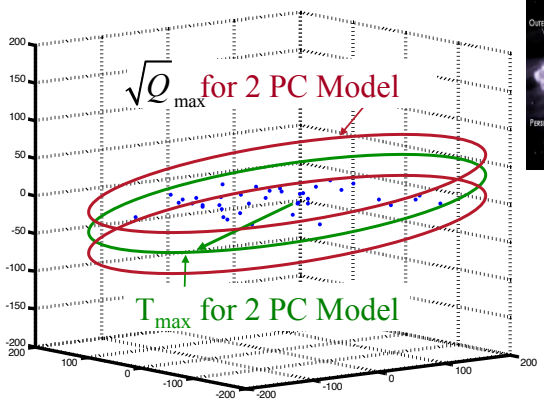
What If the Eigenvalue Plot Looks Like This:

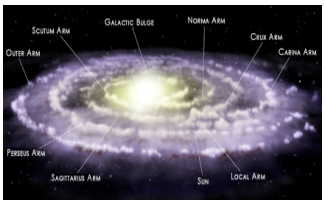


126

EIGENVECTOR RESEARCH INCORPORATED

Use Two PC Model





How many PCs would you use to describe the distribution of stars in the Milky Way?

127

EIGENVECTOR RESEARCH INCORPORATED

An Additional Method to Determine the Number of PCs to Keep

Cross Validation

What is the best method to study for a chemistry exam?

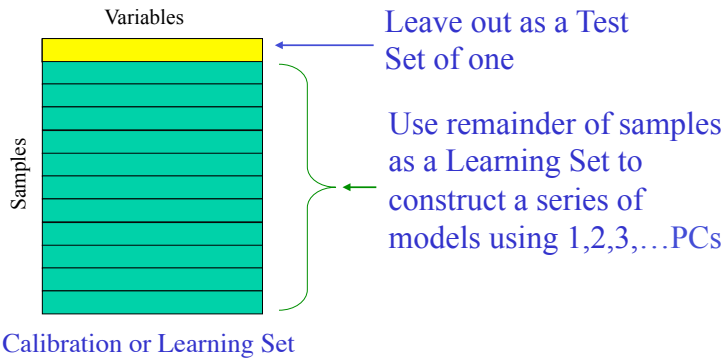
Use homework problems to make up practice exams

128

EIGENVECTOR RESEARCH INCORPORATED

Many Patterns of Cross-Validation

- Leave-One-Out Cross-Validation

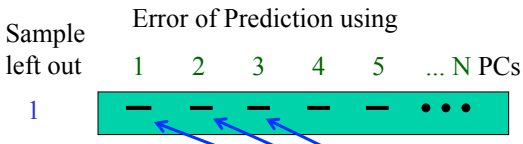


129

EIGENVECTOR RESEARCH INCORPORATED

Leave One Out Cross-Validation

For each of these models uses the scores and loading to predict the values of variables (e.g. the spectra) for the sample left out.



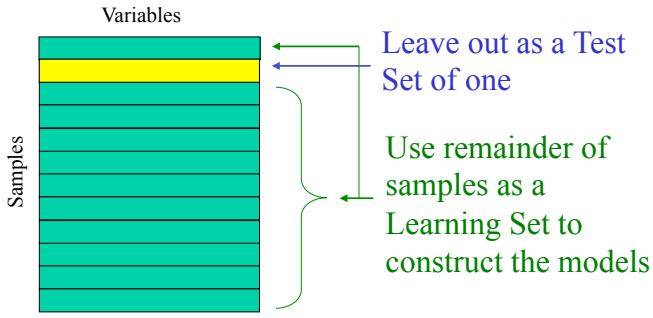
$$Error^2 = \| X_{observed} - X_{pred}(\text{using } 1,2,3\ldots PC\text{'s}) \|^2$$

130

EIGENVECTOR RESEARCH INCORPORATED

Many Patterns of Cross-Validation

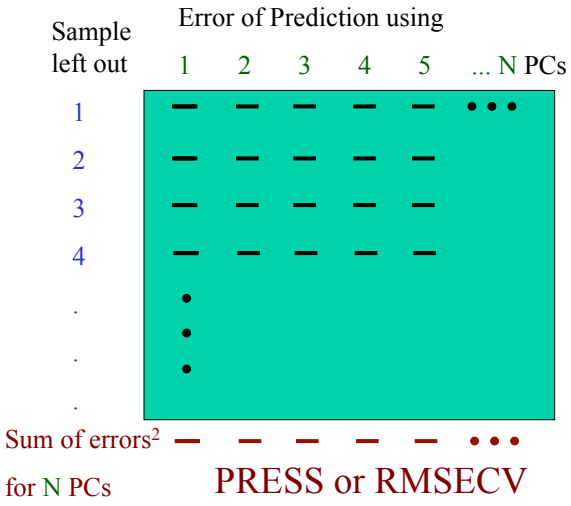
- Leave-One-Out Cross-Validation



131

EIGENVECTOR RESEARCH INCORPORATED

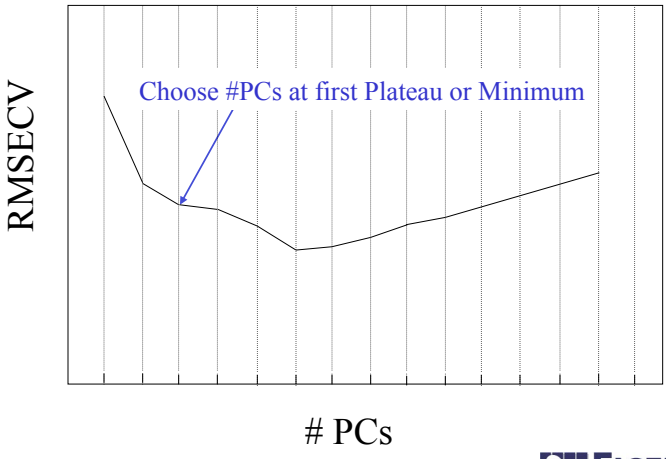
Leave One Out Cross-Validation



132

EIGENVECTOR RESEARCH INCORPORATED

Construct PRESS Plot

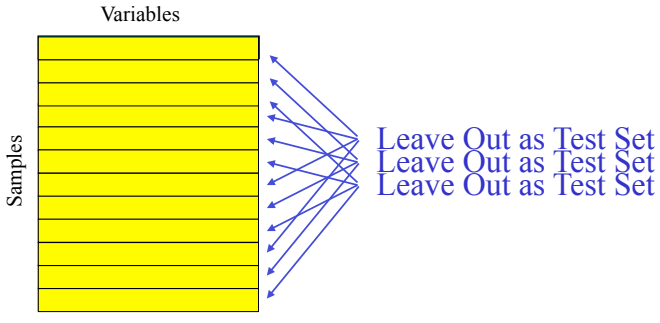


133

EIGENVECTOR RESEARCH INCORPORATED

Many Patterns of Cross-Validation

- Leave-One-Out Cross-Validation
- Venetian Blind

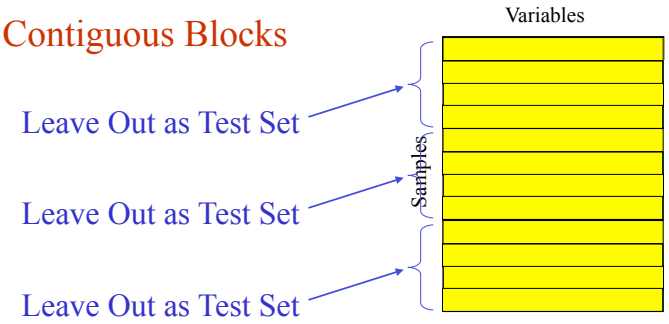


134

EIGENVECTOR RESEARCH INCORPORATED

Many Patterns of Cross-Validation

- Leave-One-Out Cross-Validation
- Venetian Blind
- Contiguous Blocks

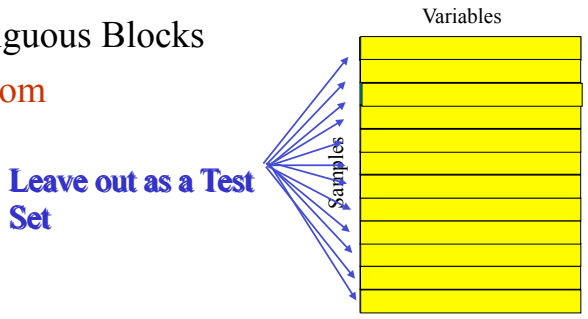


135

EIGENVECTOR RESEARCH INCORPORATED

Many Patterns of Cross-Validation

- Leave-One-Out Cross-Validation
- Venetian Blind
- Contiguous Blocks
- Random



136

EIGENVECTOR RESEARCH INCORPORATED

Which CV method do I use?

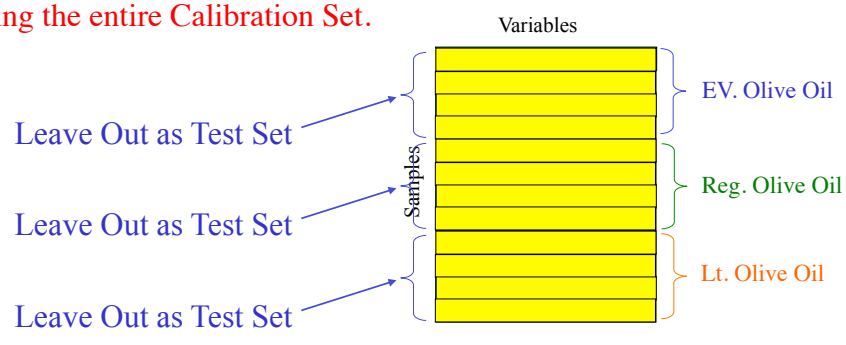
- Use Leave-one-out CV for fewer than about 16 samples in the Calibration (learning) Set.
- Your choice for larger Calibration Sets.
- Number of Splits = approximately the square root of the number of samples (24 sample, use 5 splits)
  - $1/5^{th}$ of the data is left out at a time
- Make sure that a representative set is always left behind to construct good models.

137

EIGENVECTOR RESEARCH INCORPORATED

For example, if you chose Contiguous Blocks for Olive Oil Model

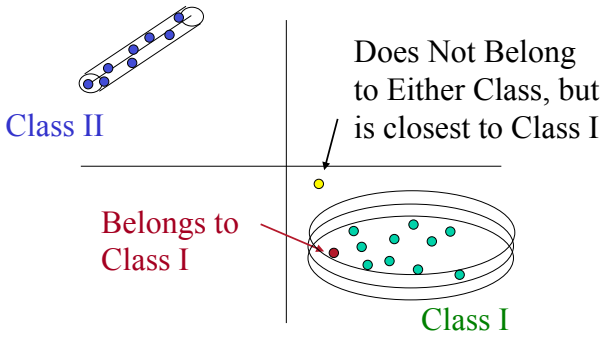
Once the number of PCs has been determined, the model is created using the entire Calibration Set.



138

EIGENVECTOR RESEARCH INCORPORATED

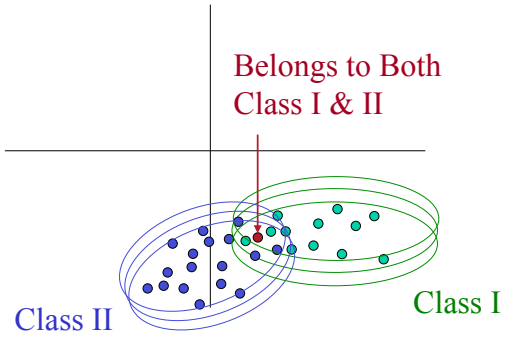
SIMCA tells if a Sample falls within the Boundaries of a Class



139

EIGENVECTOR RESEARCH INCORPORATED

With SIMCA - A Sample May Belong to Two or More Classes at the Same Time



140

EIGENVECTOR RESEARCH INCORPORATED

You Can Change the Rules for Failing to Be in the Group

- Outside T only
- Outside Q only
- Either Outside T or Outside Q
- Outside some combined limit, *e.g.* $\sqrt{(T^2 + Q^2)}$

These pictures have been using this rule

141

EIGENVECTOR RESEARCH INCORPORATED

According to these Rules SIMCA is the Only Modern Classification Method that Allows a Sample to Belong to More than One Class or No Class at All

But Rules are Made to Be Broken

Or at least changed according to our needs

142

EIGENVECTOR RESEARCH INCORPORATED

Sample Classification Predictions

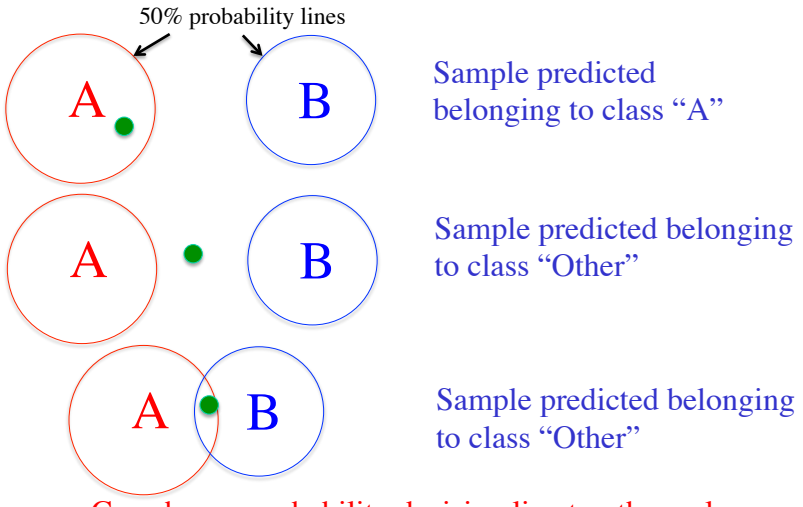
- Class Prediction Member by Class
  - Class Prediction Member – each class
  - Class Prediction Member – unassigned
  - Class Prediction Member – multiple

  } Traditional SIMCA Rules
- Class Prediction Strict
- Class Prediction Most Probable
- Class Prediction Probability by Class

143

EIGENVECTOR RESEARCH INCORPORATED

Class Prediction Strict

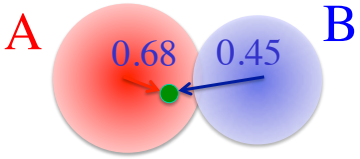


Can change probability decision line to other values

144

EIGENVECTOR RESEARCH INCORPORATED

Class Prediction Most Probable



Sample predicted more probable belonging to class "A" than to class "B"

A "soft class" model

Prob A=90% & Prob B=89%, Then assigned to Class "A"

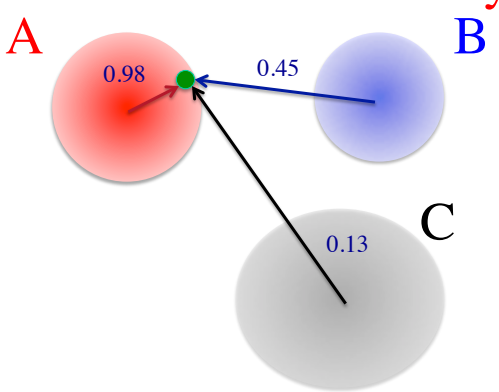
Prob A=0.02% & Prob B=0.01%, Then assigned to Class "A"

"Multiple" Class and "Unassigned" Class assignments not allowed

145

EIGENVECTOR RESEARCH INCORPORATED

Class Prediction Probability by Class



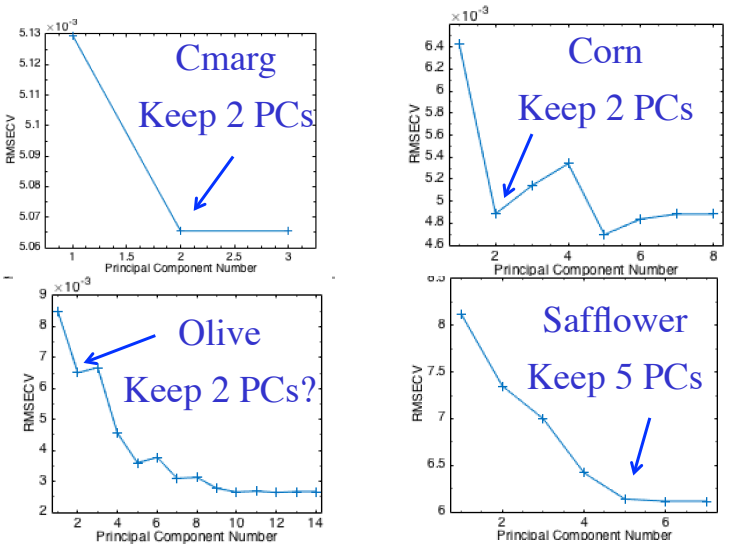
Probability calculated for each class.

Useful when you need to report a confidence of assignment or need to derive special rules for class assignment.

146

EIGENVECTOR RESEARCH INCORPORATED

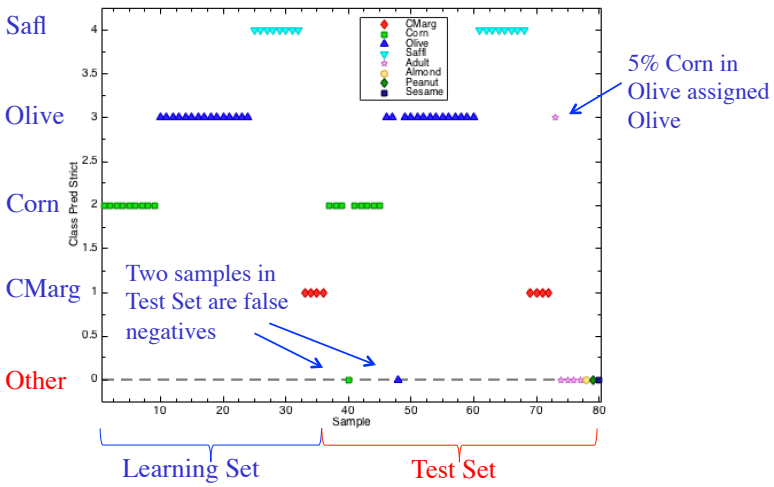
PRESS Plots for Each Class



147

EIGENVECTOR RESEARCH INCORPORATED

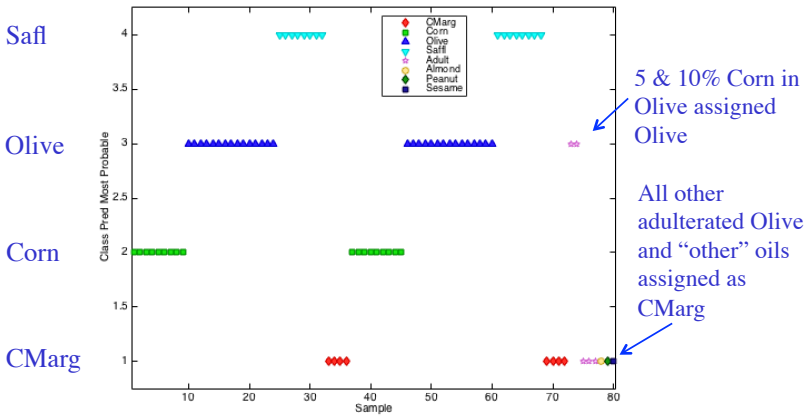
Strict Prediction



148

EIGENVECTOR RESEARCH INCORPORATED

Most Probable

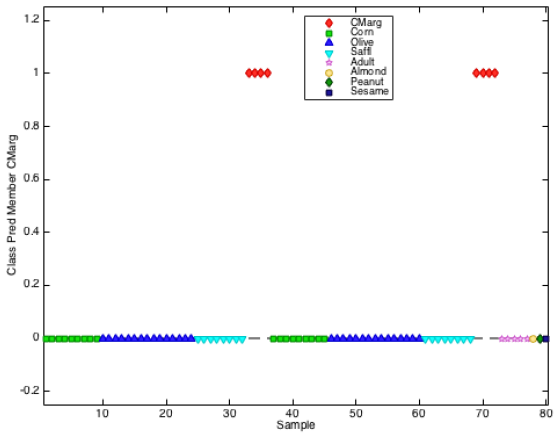


Why is this not useful for this data?

149

EIGENVECTOR RESEARCH INCORPORATED

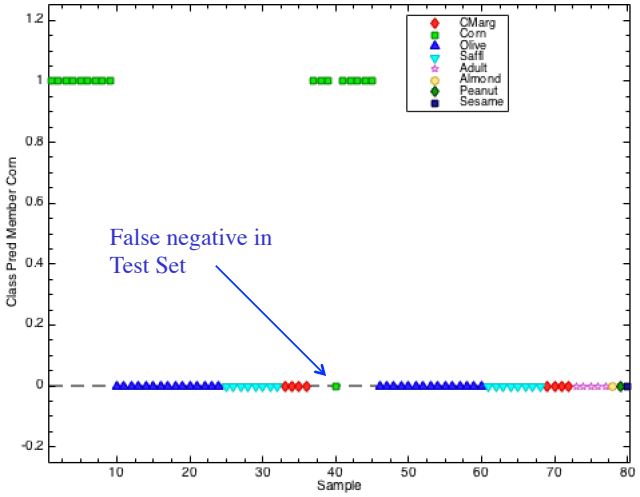
Class Predicted Member CMarg



150

EIGENVECTOR RESEARCH INCORPORATED

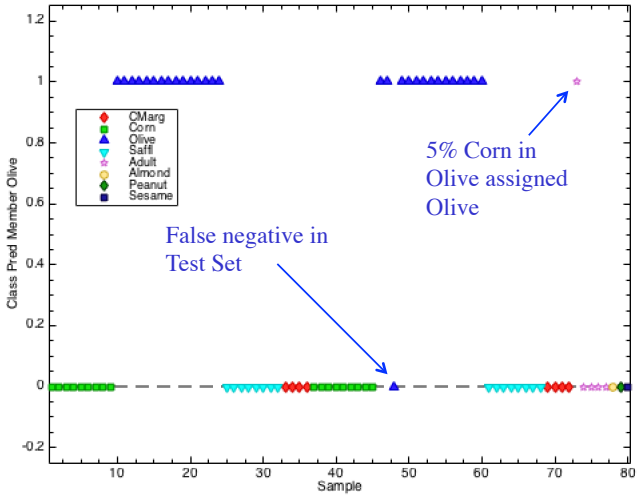
Class Predicted Member Corn



151

EIGENVECTOR RESEARCH INCORPORATED

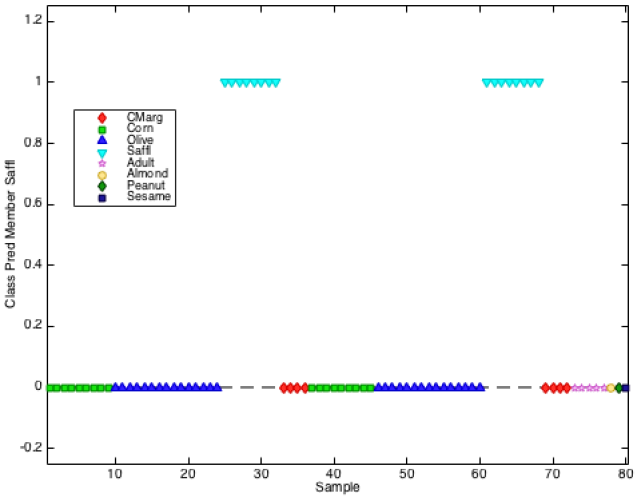
Class Predicted Member Olive



152

EIGENVECTOR RESEARCH INCORPORATED

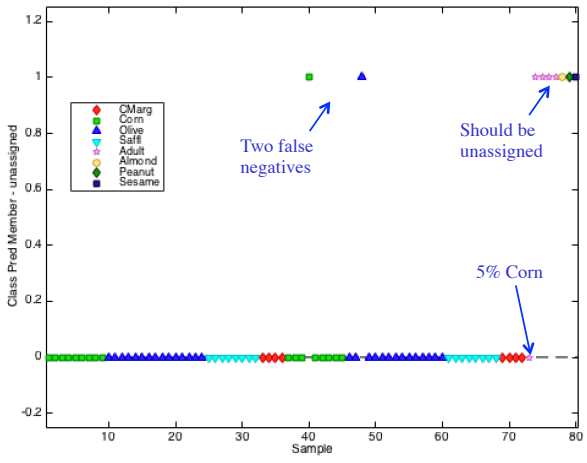
Class Predicted Member Safflower



153

EIGENVECTOR RESEARCH INCORPORATED

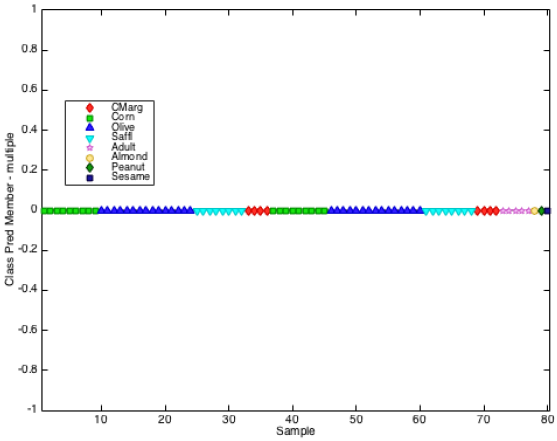
Class Predicted Member Unassigned



154

EIGENVECTOR RESEARCH INCORPORATED

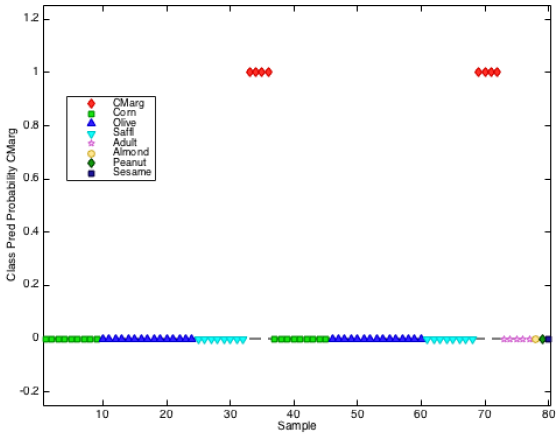
Class Predicted Member Multiple



155

EIGENVECTOR RESEARCH INCORPORATED

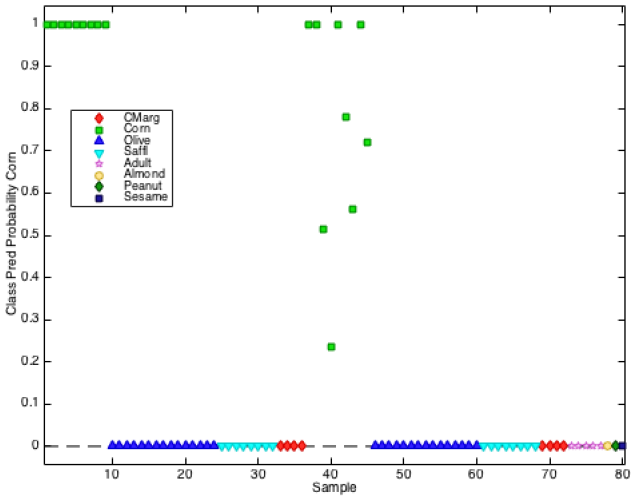
Class Predicted Probability CMarg



156

EIGENVECTOR RESEARCH INCORPORATED

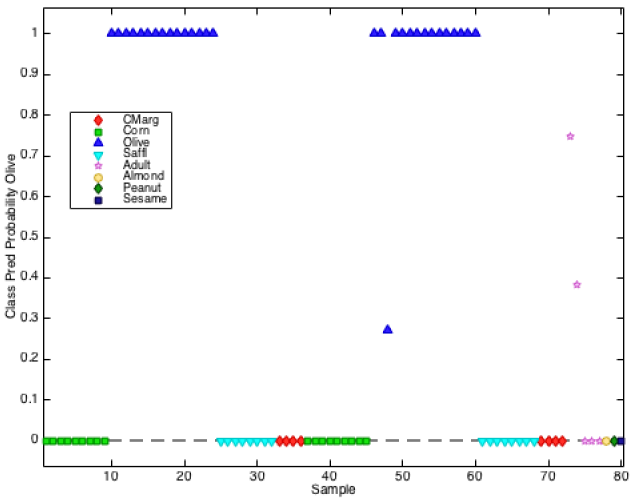
Class Predicted Probability Corn



157

EIGENVECTOR RESEARCH INCORPORATED

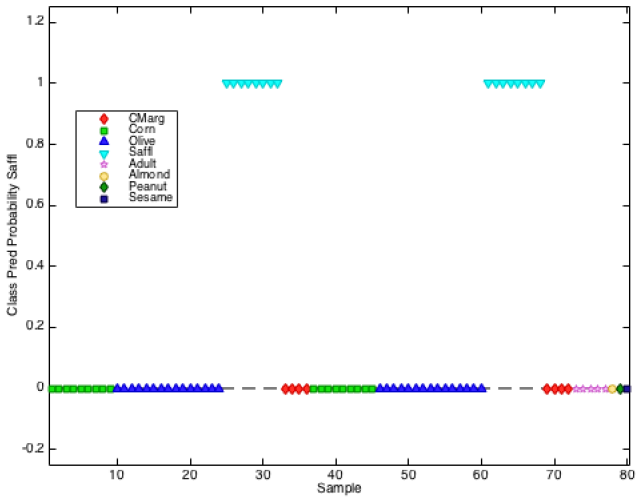
Class Predicted Probability Olive



158

EIGENVECTOR RESEARCH INCORPORATED

Class Predicted Probability Safflower



159

EIGENVECTOR RESEARCH INCORPORATED

Important Notice: Before Applying SIMCA Models to Real Unknowns

- Play with the number of PCs in the models to see if you can decrease False Negatives without increasing False Positives.
- Determine if you would rather live with false negatives or false positives.
- Validate Thoroughly With a Well Designed, Independent Test Set!
- Models Do Not Last Forever.
- Re-validate Often and Rebuild If Necessary.

160

EIGENVECTOR RESEARCH INCORPORATED

Outline (Part 2)

- Regression Motivation & Rational
  - What is Regression?
  - Why use Regression?
- Classical Least Squares (CLS)
- Multiple Linear Regression (MLR)
- Principal Components Regression (PCR)
- Partial Least Squares Regression (PLS)
- Summary

EIGENVECTOR RESEARCH INCORPORATED

We Can't Always Measure What We Want

- Often measurements must be made on something else and the property of interest must be inferred from these measurements.
- This is the idea behind inferential sensing where we measure variables that are available in a timely manner to predict something that is more difficult (or more expensive) to obtain.

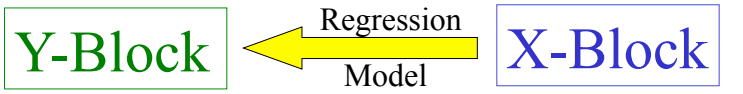
EIGENVECTOR RESEARCH INCORPORATED

What's Measured:

- Volume of $AgNO_3$ titrant to get moles of chloride in the sample.
- Absorbance at 254 nm to get the concentration of benzene in solution.
- NIR spectrum of gasoline to yield octane number.

EIGENVECTOR RESEARCH INCORPORATED

Regression



What's desired

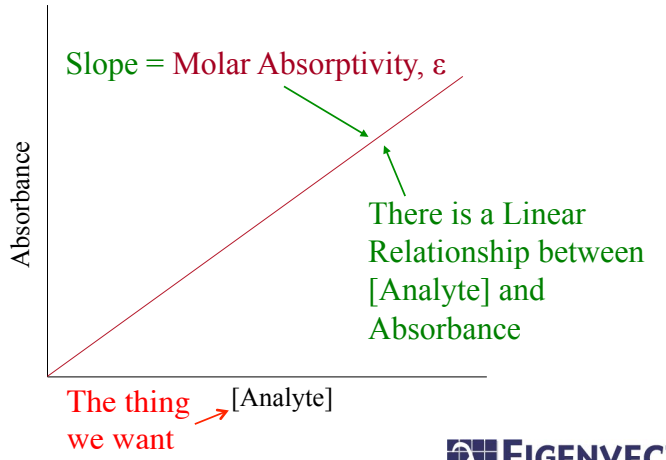
What's measured

Regression analysis identifies the dependency between two blocks of data.

Regression models are often used to obtain estimates (or predictions) for one block of data from the other.

EIGENVECTOR RESEARCH INCORPORATED

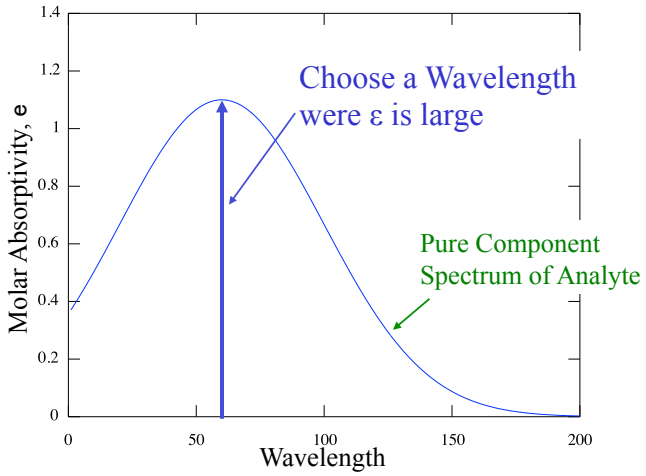
One of the Most Familiar uses of Linear Regression is from the Beer-Lambert Law



165

EIGENVECTOR RESEARCH INCORPORATED

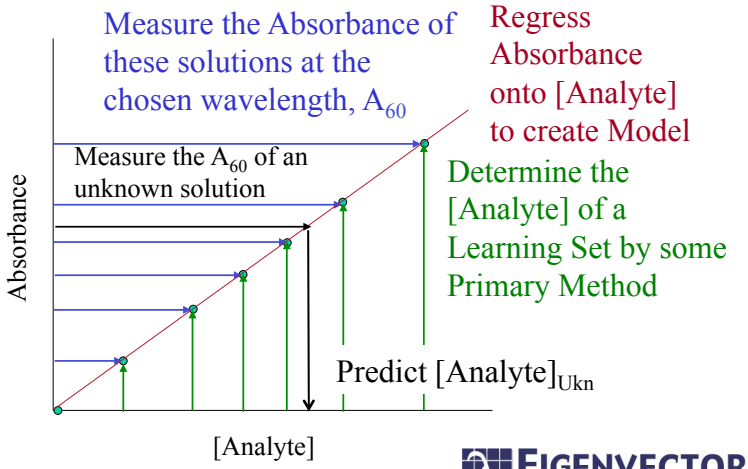
How to Create Beer-Lambert Law Model



166

EIGENVECTOR RESEARCH INCORPORATED

Create Beer-Lambert Law Model



167

EIGENVECTOR RESEARCH INCORPORATED

Outline

- Regression Motivation & Rational
- Classical Least Squares (CLS)
- Multiple Linear Regression (MLR)
- Principal Components Regression (PCR)
- Partial Least Squares Regression (PLS)
- Summary

168

EIGENVECTOR RESEARCH INCORPORATED

Sometimes Must Look at an Equation

Classical Least Squares Model = CLS:

$$A_{60} = [Analyte]\varepsilon + Error$$

Dependent Variable

Independent Variable

Regression Coefficient

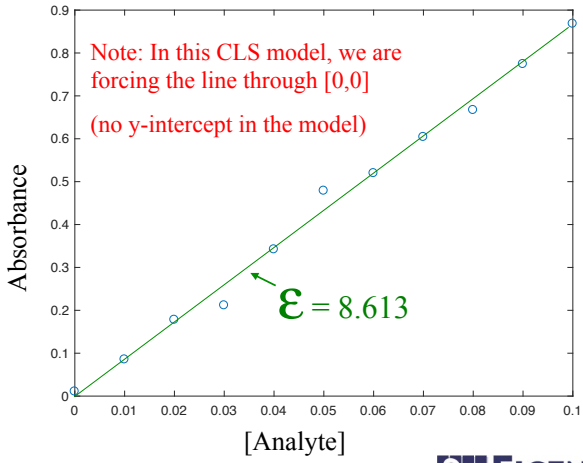
Error in measuring A

Notice the thing we ultimately want is on the right side of the equal sign.

169

EIGENVECTOR RESEARCH INCORPORATED

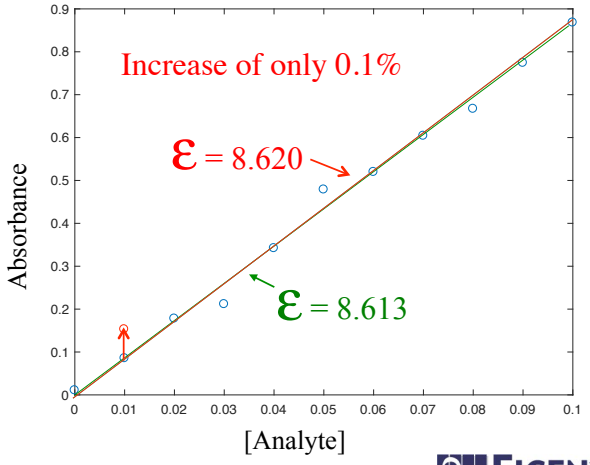
Not All Samples are Equal



170

EIGENVECTOR RESEARCH INCORPORATED

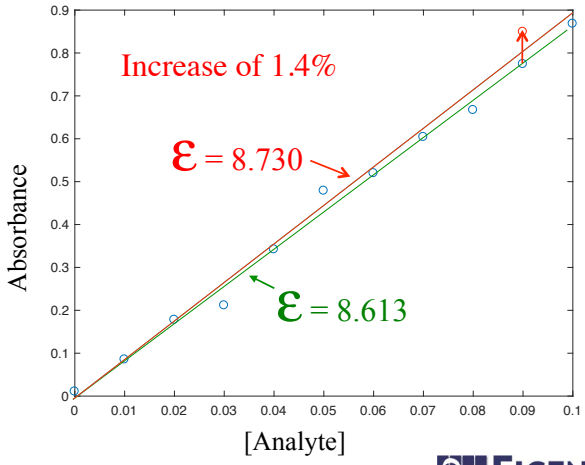
What Happens to the Slope if a Point is Moved Due to Measurement Error?



171

EIGENVECTOR RESEARCH INCORPORATED

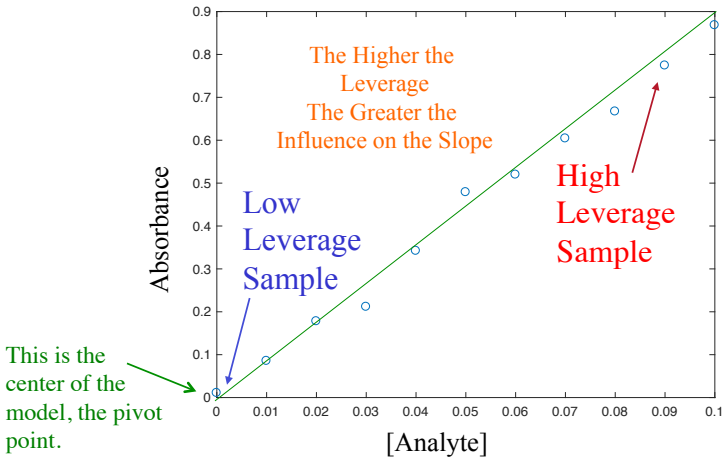
What Happens to the Slope if a Point is Moved Due to Measurement Error?



172

EIGENVECTOR RESEARCH INCORPORATED

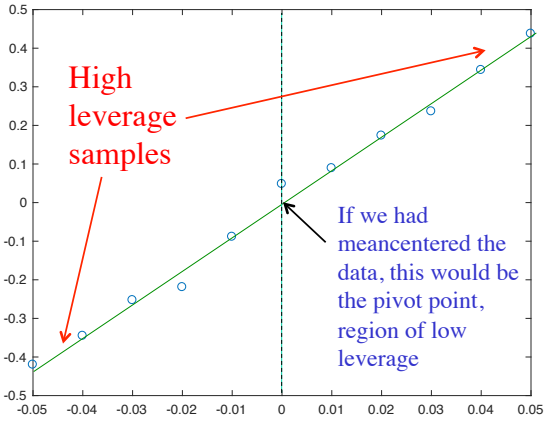
It's About Sample Leverage



173

EIGENVECTOR RESEARCH INCORPORATED

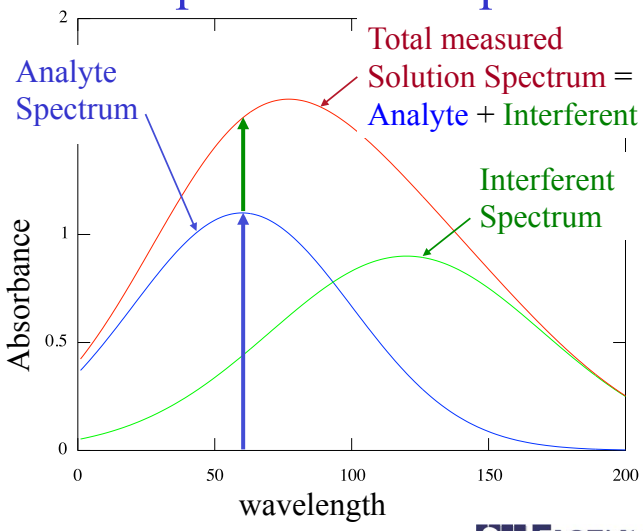
It is Important to Know the Location of the Center of the Model



174

EIGENVECTOR RESEARCH INCORPORATED

Spectral Overlap



175

EIGENVECTOR RESEARCH INCORPORATED

Beer-Lambert Law Now Looks Like:

$$A_{60} = [\text{Analyte}]\,\varepsilon_{\text{Analyte}} + [\text{Interf.}]\,\varepsilon_{\text{Intrf}}$$

Have two Concentrations to worry about

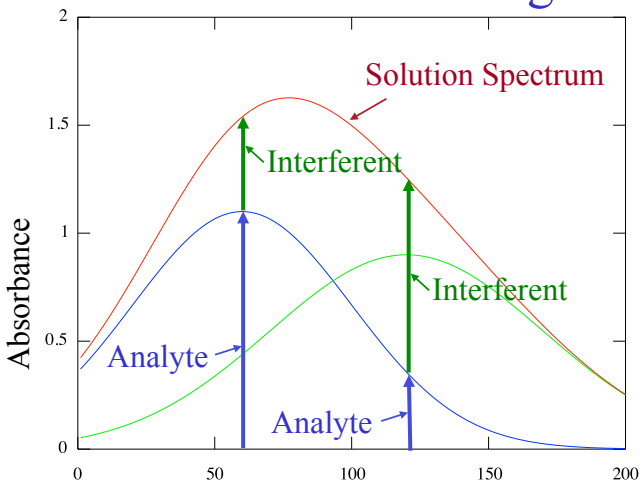
Have two Molar Absorptivities to Determine

Two Unknown Constants will require at least two equations

176

EIGENVECTOR RESEARCH INCORPORATED

What to Do?
Choose a 2nd Wavelength



177

EIGENVECTOR RESEARCH INCORPORATED

Now Have Two Equations

$$A_{60} = [\text{Analyte}]\,\varepsilon_{\text{Analyte},60} + [\text{Interf.}]\,\varepsilon_{\text{Intrf},60}$$

$$A_{120} = [\text{Analyte}]\,\varepsilon_{\text{Analyte},120} + [\text{Interf.}]\,\varepsilon_{\text{Intrf},120}$$

And FOUR Unknowns to Determine

178

EIGENVECTOR RESEARCH INCORPORATED

What to Do?

$$A_{60,I} = [\text{Analyte}]_I \, \varepsilon_{\text{Analyte},60} + [\text{Interf.}]_I \, \varepsilon_{\text{Intrf},60}$$

$$A_{120,I} = [\text{Analyte}]_I \, \varepsilon_{\text{Analyte},120} + [\text{Interf.}]_I \, \varepsilon_{\text{Intrf},120}$$

Make up a 2nd Solution using different concentrations:

$$A_{60,II} = [\text{Analyte}]_{II} \, \varepsilon_{\text{Analyte},60} + [\text{Interf.}]_{II} \, \varepsilon_{\text{Intrf},60}$$

$$A_{120,II} = [\text{Analyte}]_{II} \, \varepsilon_{\text{Analyte},120} + [\text{Interf.}]_{II} \, \varepsilon_{\text{Intrf},120}$$

Solve the equations simultaneously for the four εs

179

EIGENVECTOR RESEARCH INCORPORATED

Hey, I Thought This Workshop Was without Equations!?

180

EIGENVECTOR RESEARCH INCORPORATED

Don't Worry About It
The computer will do the work

Example:

- Make up two solutions of known [Analyte] & [Interferent] (four known concentrations).
- Measure Absorbance of both solutions at two wavelengths.
- CLS will give you the four Molar Absorptivities.
- Measure Absorbance of unknown solution at two wavelengths.
- CLS will estimate concentrations of both species/analytes in the unknown.

EIGENVECTOR RESEARCH INCORPORATED

Which Wavelengths?

- Can use (almost) all of them.
  - don't use wavelengths that can't be trusted
    - *e.g.* out of dynamic range or pure noise
- CLS is a Full Spectrum Technique
  - Multi-channel Advantage: Redundant information of many wavelengths averages out much of the noise
  - Using all channels leads to estimates of "pure component spectra" of all the chromophores
    - No extra charge!

EIGENVECTOR RESEARCH INCORPORATED

Which Samples?

- Experimental Design
  - very useful, many books on the topic
- Don't vary concentrations in the same manner
  - CLS won't work if the concentrations are correlated
- Noise can be reduced by using more than the minimum number of samples in the Learning Set
  - perhaps 5 samples per chromophore

183

EIGENVECTOR RESEARCH INCORPORATED

Sixteen Solution Learning Set
Three Component System

Component:	[A]	[B]	[C]
1	0.2845	0.5406	0.1749
2	0.4560	0.3484	0.1956
3	0.3858	0.5355	0.0786
4	0.5847	0.3213	0.0940
5	0.3109	0.2591	0.4299
6	0.2501	0.3805	0.3694
7	0.4772	0.5151	0.0076
8	0.5142	0.3508	0.1350
•	•	•	•
•	•	•	•
•	•	•	•
16	0.3830	0.5563	0.0608

184

EIGENVECTOR RESEARCH INCORPORATED

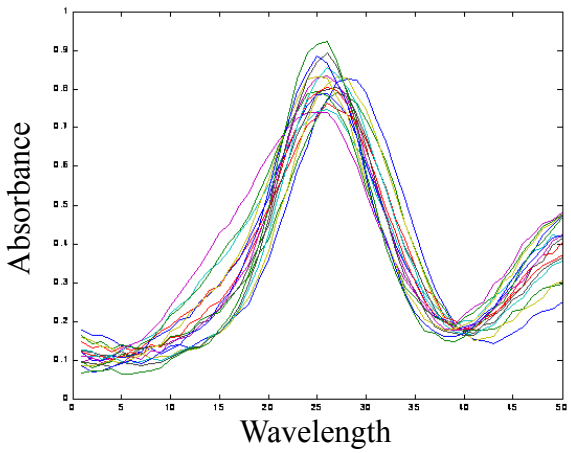
Sample Dilemma

- You want samples to cover the widest area of sample space in order to have many high leverage samples (far from the mean).
- You do not want to cover too much sample space in order not to deviate from a linear model (*e.g.* deviate from Beer's Law).
- Out of spec samples often have to be produced in the pilot plant or laboratory and may contain artifacts not present in plant samples.

185

EIGENVECTOR RESEARCH INCORPORATED

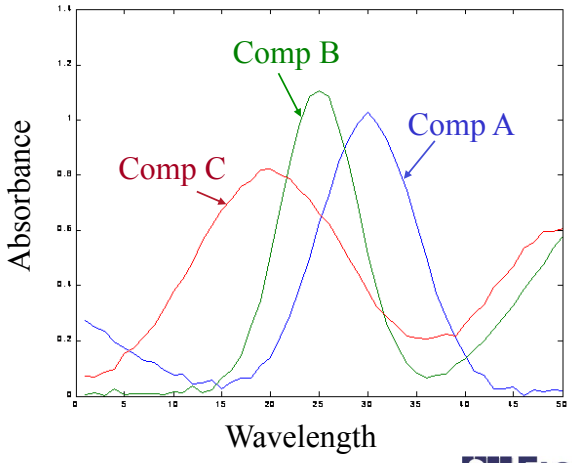
Spectra of 16 Learning Set Spectra



186

EIGENVECTOR RESEARCH INCORPORATED

Pure Component Spectra Obtained from CLS

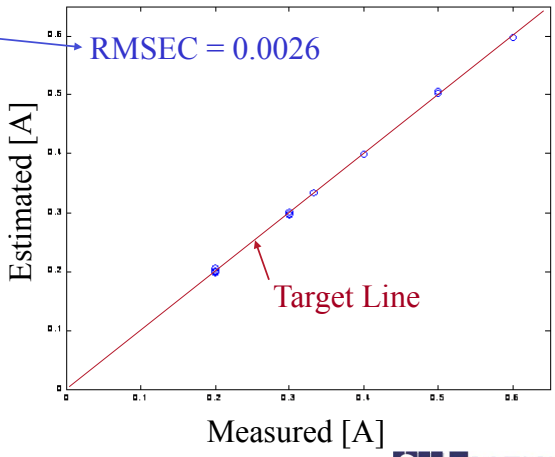


187

EIGENVECTOR RESEARCH INCORPORATED

Estimated vs. Measured [A] for Learning Set

Root Mean Square Error of Calibration

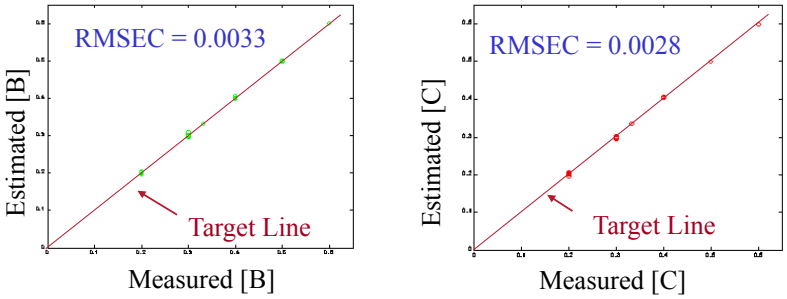


Estimated because Model learned from these samples

188

EIGENVECTOR RESEARCH INCORPORATED

Estimated vs. Measured [B] and [C] for Learning Set



189

EIGENVECTOR RESEARCH INCORPORATED

Estimation versus Prediction

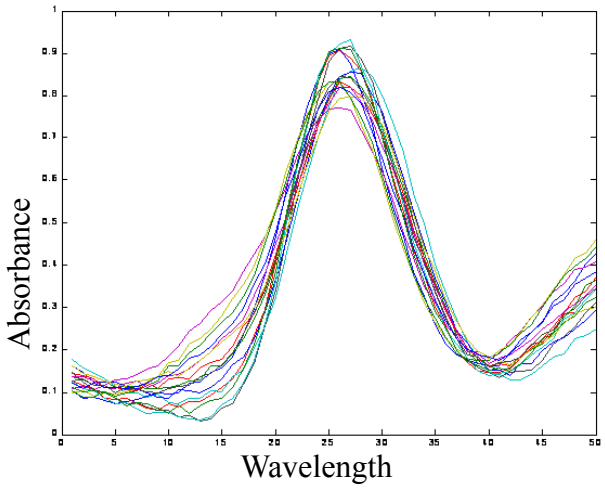
- The previous slides were for the Learning Set samples
  - This provides an estimate of the fit or calibration error
- But the real test of any model is how it does on samples it has **Never Seen Before**

→ Test Set

190

EIGENVECTOR RESEARCH INCORPORATED

Spectra of 16 Sample Test Set

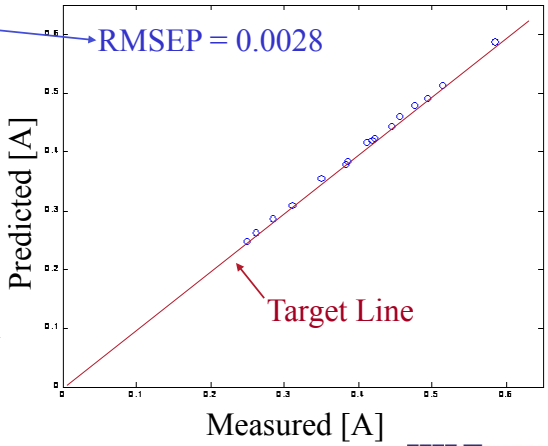


191

EIGENVECTOR RESEARCH INCORPORATED

Predicted vs. Measured [A] for Test Set

Root Mean Square Error of Prediction



Predicted because Model has never seen these samples

192

EIGENVECTOR RESEARCH INCORPORATED

Compare RMSEC with RMSEP

Component	RMSEC	RMSEP
A	0.0026	0.0028
B	0.0033	0.0027
C	0.0028	0.0031

For any given component, we want RMSEP ~ RMSEC

193

EIGENVECTOR RESEARCH INCORPORATED

Summary of CLS

Advantages

- Simultaneously determine the concentrations of many analytes in a mixture.
- Multi-channel Advantage.
- Do not need pure component spectra.
- Provides pure component spectra in the sample environment (matrix).

Disadvantages

- Need the concentration of all chromophores in each sample of the learning set.*
- Very sensitive to non-linearity.

*Not necessarily true with more advanced forms of CLS

194

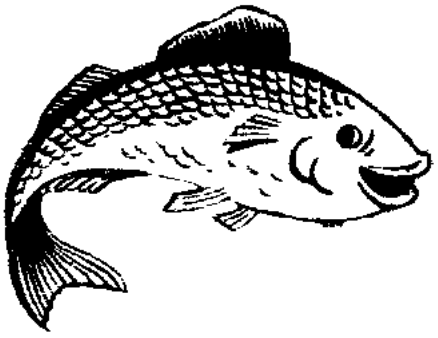
EIGENVECTOR RESEARCH INCORPORATED

Outline

- Regression Motivation & Rational
- Classical Least Squares (CLS)
- Multiple Linear Regression (MLR)
- Principal Components Regression (PCR)
- Partial Least Squares Regression (PLS)
- Summary

EIGENVECTOR RESEARCH INCORPORATED

NIR of a Fish



We do not want to know the concentration of all the components in a fish.

What all is in a fish?

EIGENVECTOR RESEARCH INCORPORATED

Small Change to the Beer-Lambert Law

Multiple Linear Regression - MLR:

$$[\text{Analyte}] = A_{60}\, b + \text{Error}$$

Dependent Variable

Independent Variable

Regression Coefficient

Error in measuring concentration

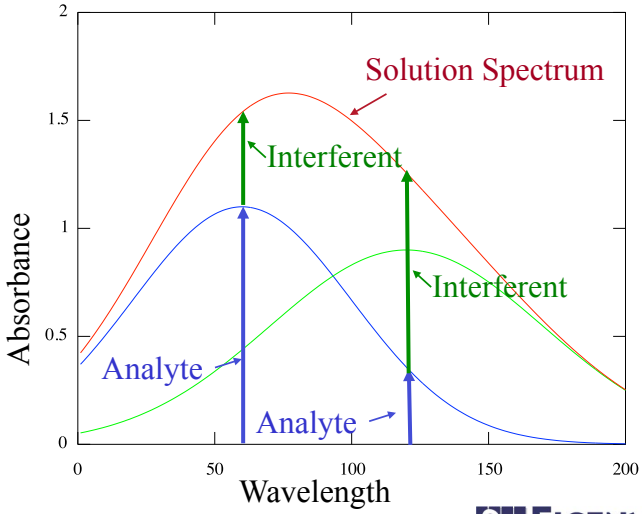
Notice the thing we ultimately want is on the left side of the equal sign.

Multiple Linear Regression (MLR) is Sometimes Referred to as Inverse Least Squares (ILS)

197

EIGENVECTOR RESEARCH INCORPORATED

Let's Look at MLR Applied to Our

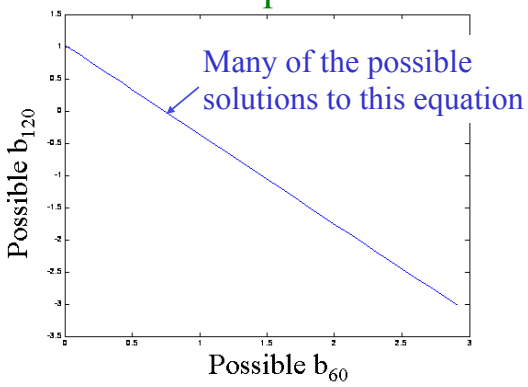


198

EIGENVECTOR RESEARCH INCORPORATED

$$[\text{Analyte}] = A_{60}\, b_{60} + A_{120}\, b_{120}$$

One Equation - Two Unknowns
No Unique Solution



199

EIGENVECTOR RESEARCH INCORPORATED

What to Do?

$$[\text{Analyte}]_{I} = A_{60,I}\, b_{60} + A_{120,I}\, b_{120}$$

Make up a 2nd Solution using different concentrations:

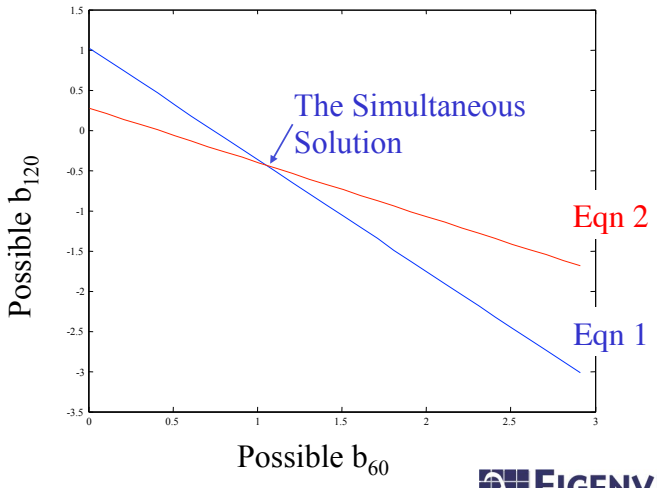
$$[\text{Analyte}]_{II} = A_{60,II}\, b_{60} + A_{120,II}\, b_{120}$$

Notice the other interfering chromophores are not in the equation.

200

EIGENVECTOR RESEARCH INCORPORATED

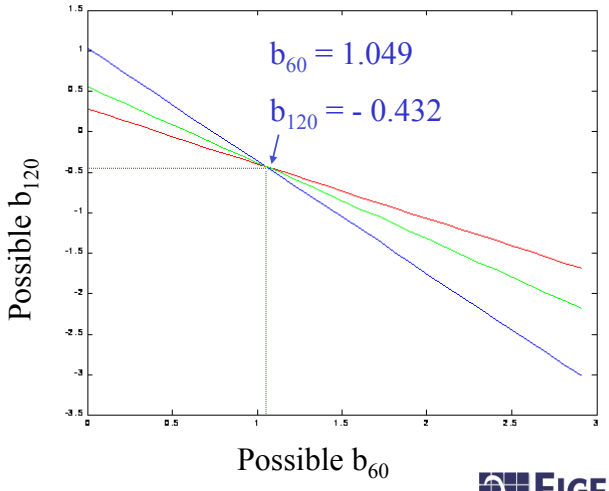
Two Equations - Two Unknowns



201

EIGENVECTOR RESEARCH INCORPORATED

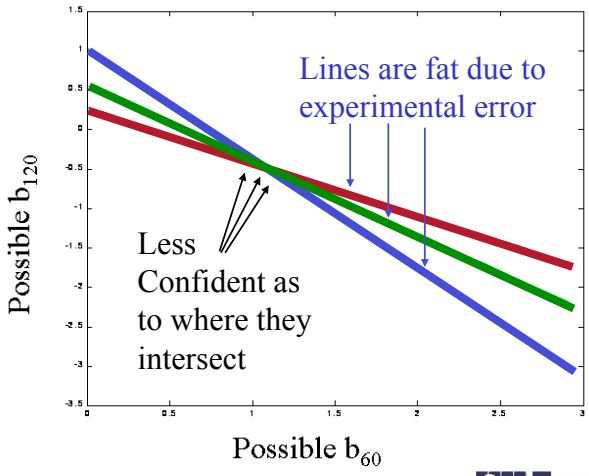
Add a Third Experiment
Overdetermined Solution



202

EIGENVECTOR RESEARCH INCORPORATED

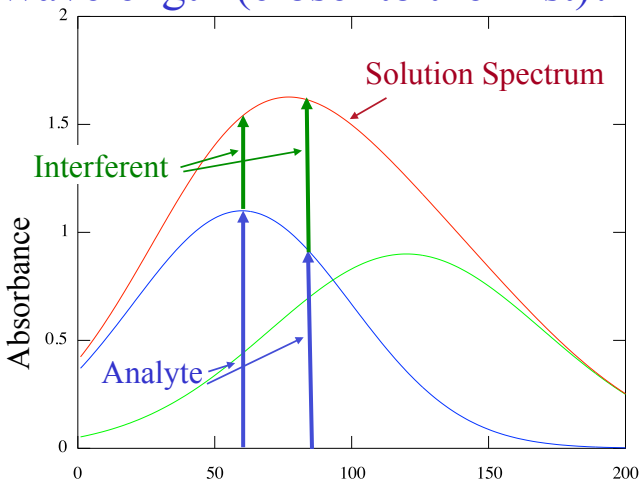
Problem Solved? Not Exactly!



203

EIGENVECTOR RESEARCH INCORPORATED

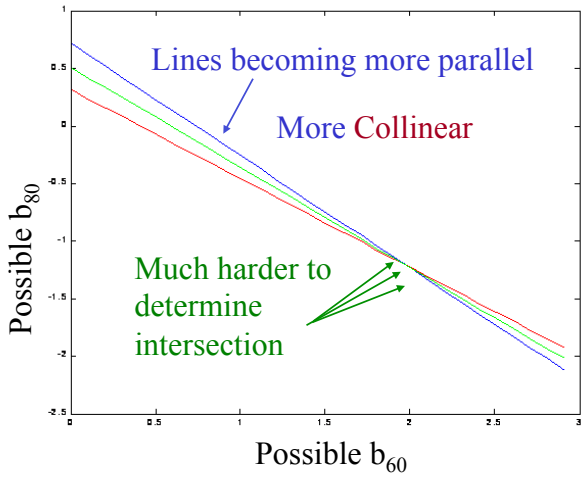
What if We Choose a Different 2nd Wavelength (closer to the first)?



204

EIGENVECTOR RESEARCH INCORPORATED

Using Wavelength 60 & 80



205

EIGENVECTOR RESEARCH INCORPORATED

Summary - MLR

Advantages

- Do not need to know the concentration of all chromophores in each sample of the learning set.
- Do not need the pure component spectra.

Disadvantages

- Best noise removal with many more samples than variables.
  - usually variables are cheap but samples are expensive
- Must use carefully chosen independent variables to avoid collinearity.
  - Lose multi-channel advantage
  - Often used just half dozen wavelengths
- Do not get pure component spectra

206

EIGENVECTOR RESEARCH INCORPORATED

Can't I have my cake and eat it too?

- *i.e.* is there a way to do MLR and keep more variables?
  - retain the multi-channel advantage
- How to handle the collinearity?
- Maybe there is a way to use PCA?

Outline

- Regression Motivation & Rational
- Classical Least Squares (CLS)
- Multiple Linear Regression (MLR)
- Principal Components Regression (PCR)
- Partial Least Squares Regression (PLS)
- Summary

EIGENVECTOR RESEARCH INCORPORATED

Problem with MLR

- How do we keep all the variables as in CLS and only have to model the analyte as in MLR?
  - *i.e.* want a full spectrum technique, but don't know the concentration of all the chromophores
- Problem with MLR is: Collinear Variables
  - Also known as: Correlated, Redundant, Parallel
- Opposite of Collinear is Orthogonal
- How do we get Orthogonal Variables?
  - Also known as: Uncorrelated, Independent, Perpendicular

EIGENVECTOR RESEARCH INCORPORATED

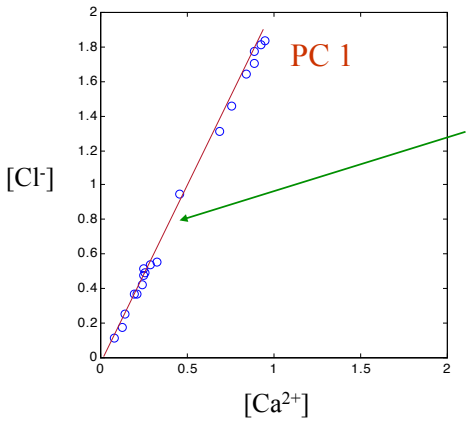
PCA

- Recall that PCA converted real variables into new orthogonal variables: Principal Components
- Loadings for each PC are orthogonal to the loadings from other PCs
  - Loadings ←→ Variables
- Scores are the positions of the samples in the new PC space.
  - Scores ←→ Samples
- Perhaps MLR on the PCA scores?

210

EIGENVECTOR RESEARCH INCORPORATED

Recall the Solutions Analyzed for Ca^{2+} and Cl^-?



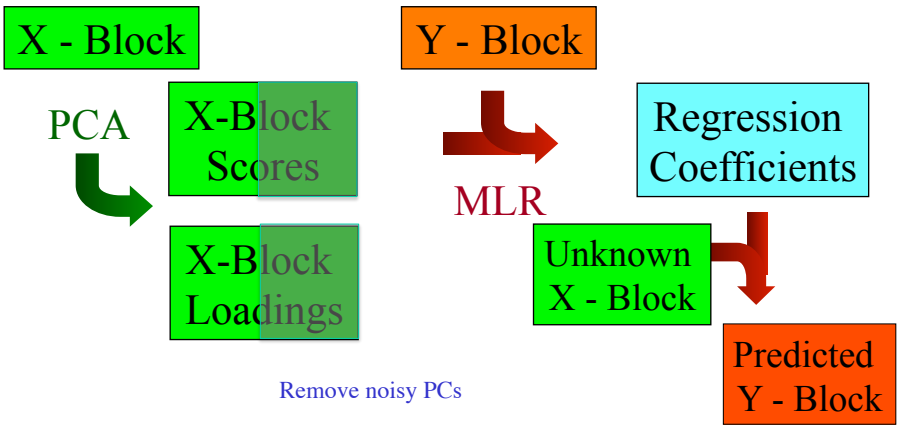
PC 1 was interpreted as related to the $[CaCl_2]$

211

EIGENVECTOR RESEARCH INCORPORATED

Principal Components Regression

PCR

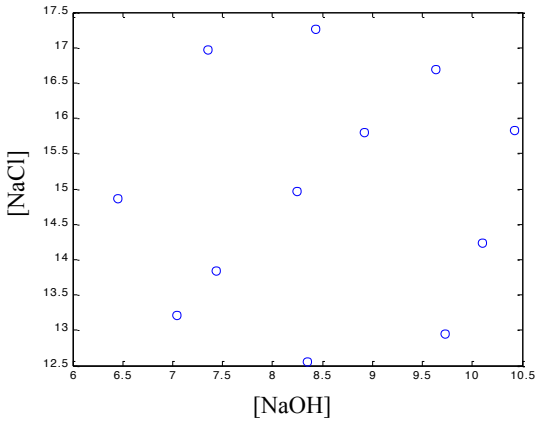


212

EIGENVECTOR RESEARCH INCORPORATED

Want to determine the concentration of NaOH in aqueous caustic brine solutions using SW-NIR

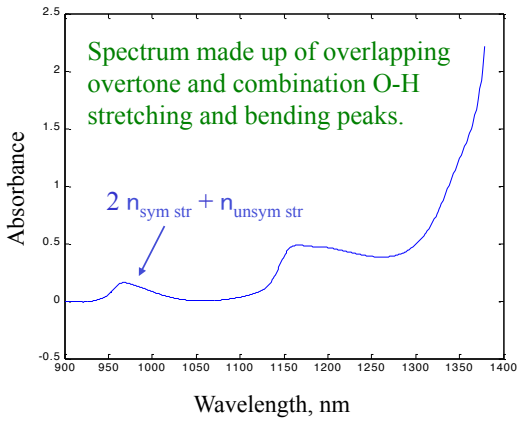
Have 12 Solutions of NaOH and NaCl in Water



213

EIGENVECTOR RESEARCH INCORPORATED

Typical SW-NIR Spectrum of Caustic Brine Solution



214

EIGENVECTOR RESEARCH INCORPORATED

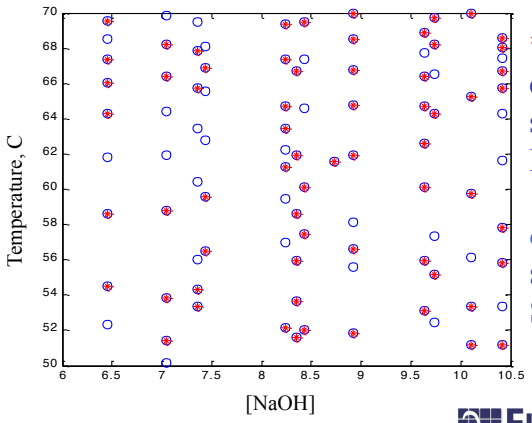
Peak Shift

- The NIR water peaks shift with changes in
  - NaCl and NaOH, and
  - temperature
- Since the temperature will vary in the application, temperature variation must be included in the Learning Set
  - although temperature need not be known to calibrate for NaOH, it must vary in the Learning Set for the model to be robust to changes in temperature.
- Interferents (e.g. NaCl & temperature) must either be kept constant forever, or varied in the learning set.

215

EIGENVECTOR RESEARCH INCORPORATED

Data: 95 Spectra SW-NIR of 12 Caustic Brine Solutions from 50-70° C



* Randomly choose 65 samples as Learning Set

o Remaining 30 samples are Test Set

Seasholtz MB., *Chemom. Intell.Lab. Syst.* 1999; **45**: 55–63.

EIGENVECTOR RESEARCH INCORPORATED

PCR on Learning Set

- Should we mean center or autoscale?
  - Usually just mean center spectra
  - recall that this means that we're allowing for an intercept other than zero
- How many PCs to keep in the model?
  - RMSEC vs #PC's
  - RMCECV vs #PC's Cross-validation

EIGENVECTOR RESEARCH INCORPORATED

Cross Validation

- Works just like it did for PCA
- Except we have two types of errors (Residuals):
  - X-block Error – same as for PCA
  - Y-block Error = $Y_{pred} - Y_{meas}$

Used to calculate RMSECV

EIGENVECTOR RESEARCH INCORPORATED

Cross-Validation Rules of Thumb

- Only use Leave-One-Out CV for fewer than 16 samples.
- Best results are obtained if you take out Square Root of Number of Samples at a time.
- "Genuine Replicates" can be split between the Learning and Test Sets
  - "'genuine replicates' are repetitions which are subject to ALL the sources of error that affect runs made at different experimental conditions"*
- If simple repeat measurements, keep them together, *i.e.* have all in either the Learning Set or Test Set.

*Box, Hunter, and Hunter, "Statistics for Experimenters", Wiley (1978)

EIGENVECTOR RESEARCH INCORPORATED

Cross-Validation for PCR Example

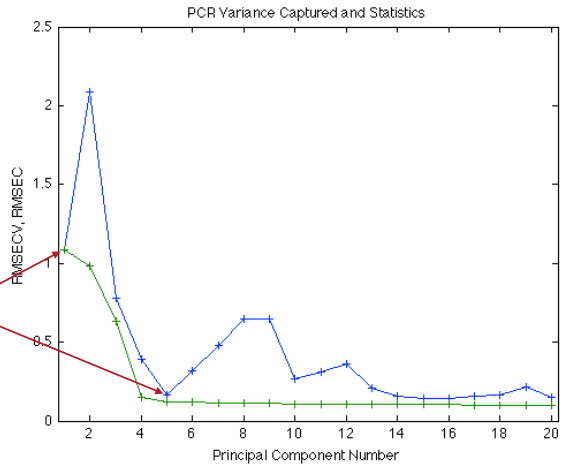
- Venetian Blind
- 8 Splits (approx. ~65 samples)
- Calculate RMSECV for up to 20 PCs
  - expect the number to be <20

220

EIGENVECTOR RESEARCH INCORPORATED

PRESS Plot

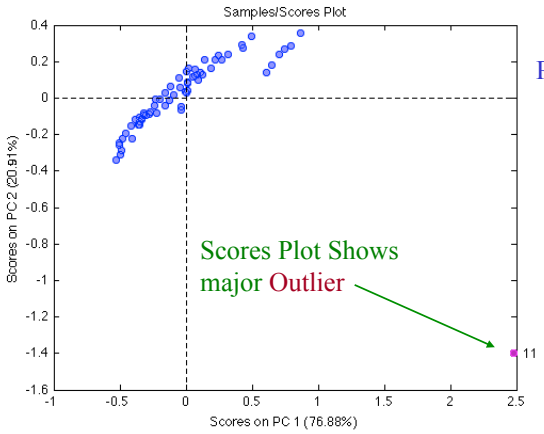
Double Minimum often suggests Outliers



221

EIGENVECTOR RESEARCH INCORPORATED

Keep 5 PCs
just to see what is going on

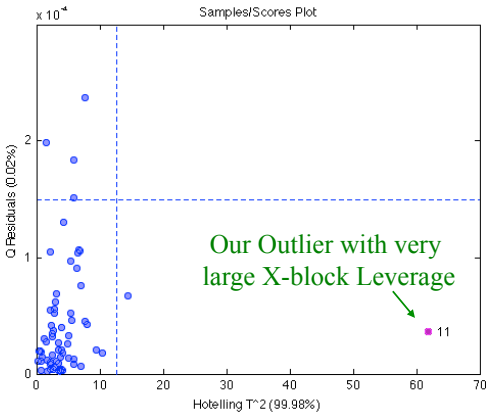


PC 1+2 ~97.79%
of the variance
in the X-block

EIGENVECTOR RESEARCH INCORPORATED

Outliers Can Also be Found by Examining the X-Block Influence Plot

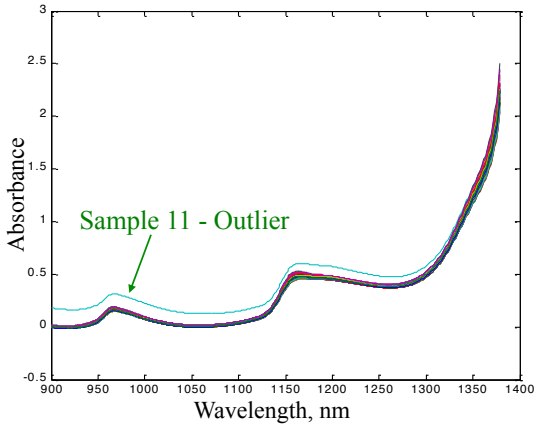
X-block Residual
- How much each
spectrum differs
from the PCA
Model (retained
PCs) of spectra.



It appears to be
within the model,
because it pulls
the model to
itself

EIGENVECTOR RESEARCH INCORPORATED

Spectra of Learning Set



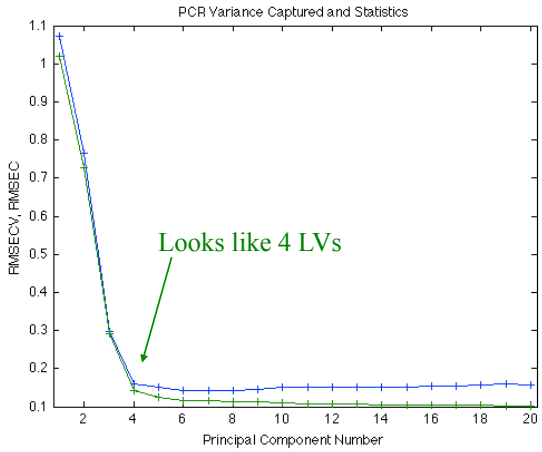
EIGENVECTOR RESEARCH INCORPORATED

Important Lessons

- Always plot spectra to see if there are any obvious outliers.
  - Plot your data
- Do PCA on your X-Block and examine the scores, Q, and T^2 for Outliers before building calibration model
  - Exploratory data analysis

EIGENVECTOR RESEARCH INCORPORATED

Delete Outlier Sample 11, Rebuild Model and Examine PRESS Plot

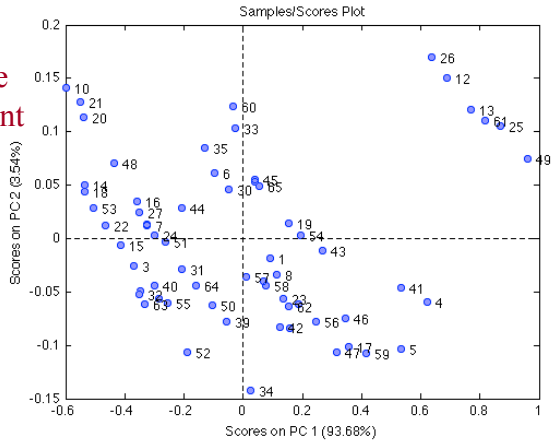


226

EIGENVECTOR RESEARCH INCORPORATED

Scores Plot

additional outliers are not apparent



Identical to PCA Scores Plot

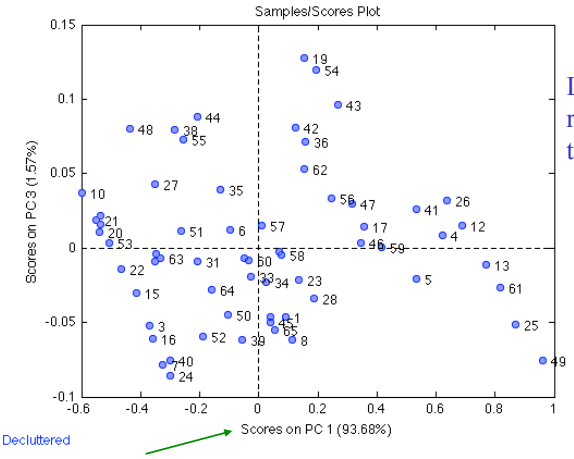
Lots of Structure to examine when we are done finding outliers.

PC 1+2 ~97.22%

227

EIGENVECTOR RESEARCH INCORPORATED

Examine All the PC Scores



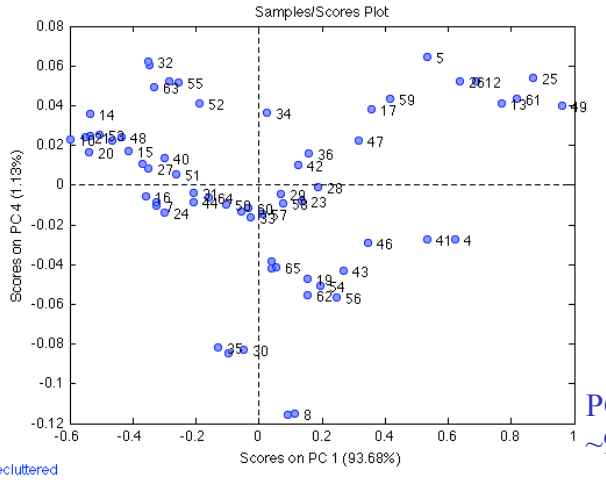
Look for samples removed from the rest.

PC 1+3
~95.25%

Keep PC 1 as the x-axis

EIGENVECTOR RESEARCH INCORPORATED

Scores PC 4 vs PC 1

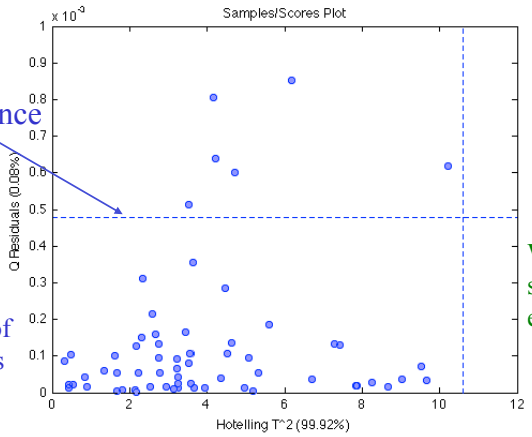


PC 1+4
~94.81%

EIGENVECTOR RESEARCH INCORPORATED

X-Block Influence Plot

95% Confidence Level



We have no samples with excessive T^2

At 95% CL we would expect 1 of every 20 samples (about 3) above line.

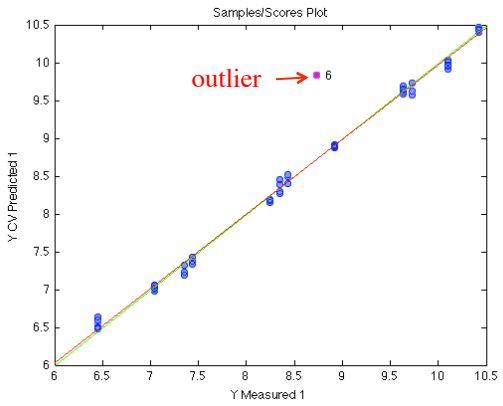
EIGENVECTOR RESEARCH INCORPORATED

Looking for X-block Outliers

- Look for isolated sample(s) in Scores Plots.
- Look for samples with excessive Q Residual or Hotelling's T^2.
- An Outlier is not necessarily a bad sample, just unique. Examine it:
  - If good, add more like it to Learning Set.
  - If bad, fix it or remove it.

EIGENVECTOR RESEARCH INCORPORATED

How Did We Do?
Estimated [NaOH] vs. Measured [NaOH]



This plot is for the Learning Set, but since we are doing CV, it is a "Predicted" Y values.

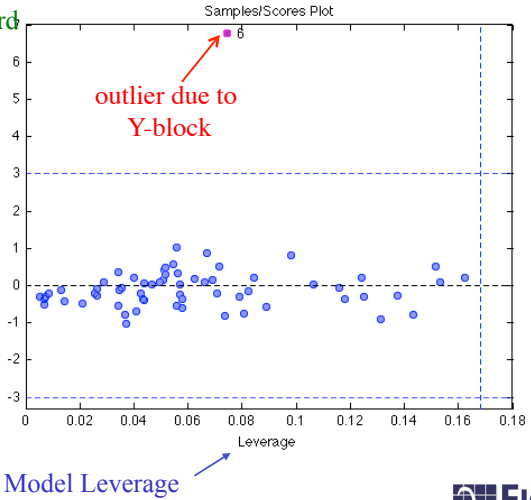
Deviations from the diagonal are used to calculate RMSECV.

232

EIGENVECTOR RESEARCH INCORPORATED

Model or Y-Block Influence Plot

Residual expressed in units of Standard Deviation



Y Residual

Large Leverage ⟶ implies error in **X**-block

Model Leverage

233

EIGENVECTOR RESEARCH INCORPORATED

Looking for Outliers

- Samples with excessive Y-block Residuals of Estimation
  - Look for problem in **Y**-block
- Samples with excessive X-block Leverage or Q (X-Block) Residual
  - Look for problem in **X**-block
- Remove or fix samples with problems.
- If sample OK, then add more like it.

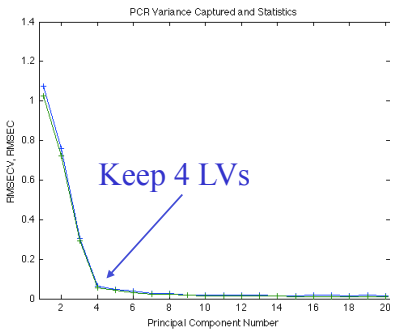
EIGENVECTOR RESEARCH INCORPORATED

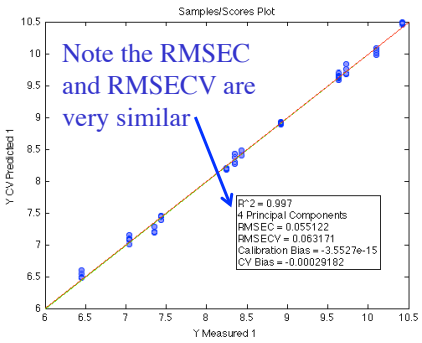
Remove or Fix Sample #6

- It turns out that there was an "entry error" for the "known" [NaOH] for Sample #6
  - The measured was 9.7325
  - Entered in the data was 8.7325
- Once corrected, we can make another model
- Removing or correcting outliers and rebuilding the model often reveals new outliers not seen before.
- Check for outlier again using Influence & Scores Plots
- Modeling tends to be iterative
  - Use what is learned at each step to help in subsequent models

EIGENVECTOR RESEARCH INCORPORATED

PRESS and Prediction Plot Plots

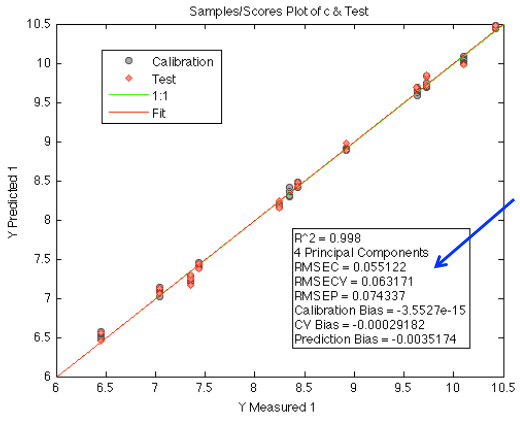




236

EIGENVECTOR RESEARCH INCORPORATED

Apply Model to Test Set



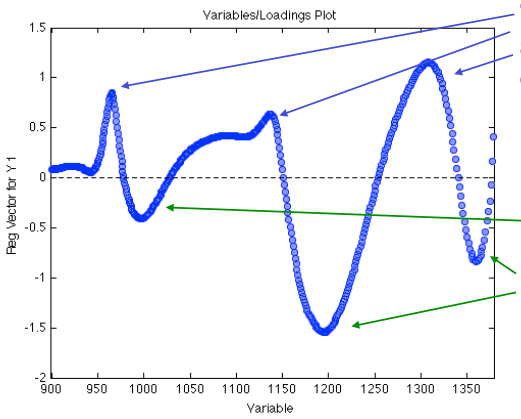
237

EIGENVECTOR RESEARCH INCORPORATED

This Is the Model Regression Coefficients

Danger:

Be careful interpreting Regression Coefficients from PCR (and PLS). They are ***not*** pure component spectra.



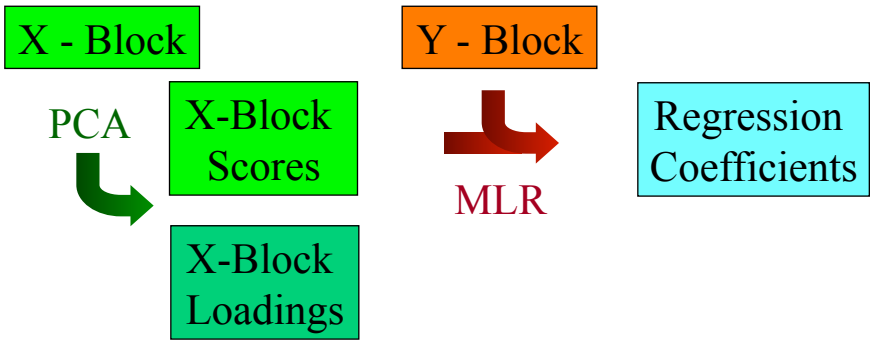
These variables are positively correlated to [NaOH]

These variables are negatively correlated to [NaOH], correcting for an overlapping interferent

238

EIGENVECTOR RESEARCH INCORPORATED

Recall How PCR Works



239

EIGENVECTOR RESEARCH INCORPORATED

Problem with PCR

- The PCs were created from the **X**-Block without any help from the **Y**-Block
  - and sorted in order of variance captured
- The PCs useful for predicting the **Y**-Block may be deep down in the pile, … with all the noise
- Shouldn't we bring in the **Y**-Block earlier in the process, at least as a consultant on the construction of the Principal Components ←→Latent Variables?

240

EIGENVECTOR RESEARCH INCORPORATED

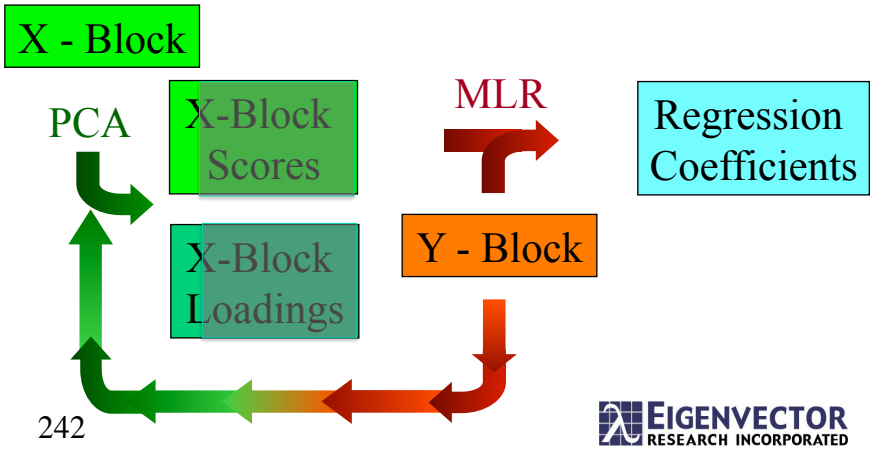
Outline

- Regression Motivation & Rational
- Classical Least Squares (CLS)
- Multiple Linear Regression (MLR)
- Principal Components Regression (PCR)
- Partial Least Squares Regression (PLS)
  - PLS-1
  - PLS-2
- Summary

241

EIGENVECTOR RESEARCH INCORPORATED

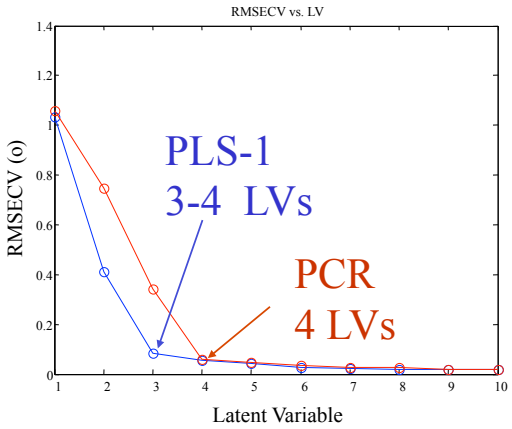
Modification to PCR
Partial Least Squares, PLS-1



242

EIGENVECTOR RESEARCH INCORPORATED

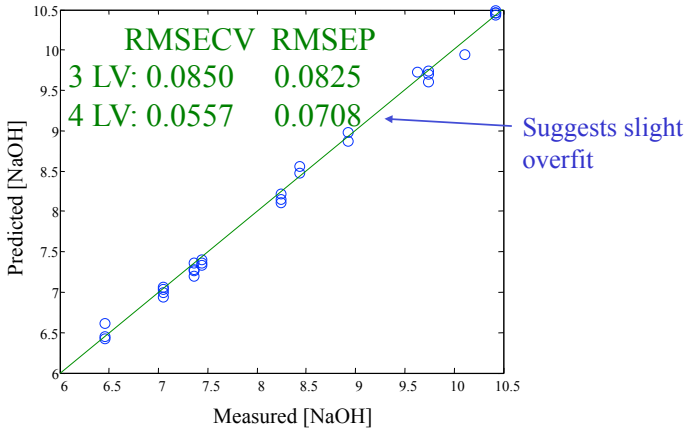
Compare PRESS Plots



243

EIGENVECTOR RESEARCH INCORPORATED

PLS-1 Prediction of Test Set with 3 LVs



244

EIGENVECTOR RESEARCH INCORPORATED

PLS-1 Compared to PCR

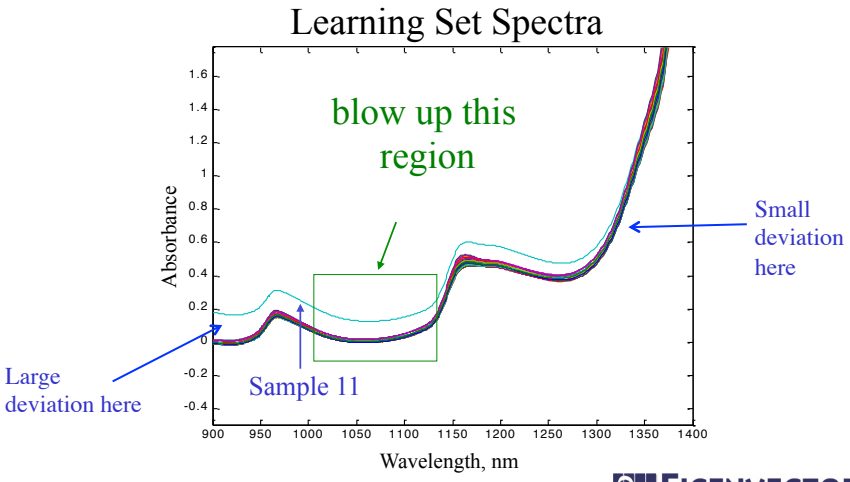
- The hope is that PLS-1 will require fewer latent variables than PCR requires principal components which suggests that there will be less noise
- The two methods often provide very similar results
- PLS usually brings more useful information into the earlier LVs for easier interpretation

Our example:	RMSECV	RMSEP
PLS-1 (3 LV)	0.0850	0.0825
PLS-1 (4 LV)	0.0557	0.0708
PCR (4 PC)	0.0592	0.0743

245

EIGENVECTOR RESEARCH INCORPORATED

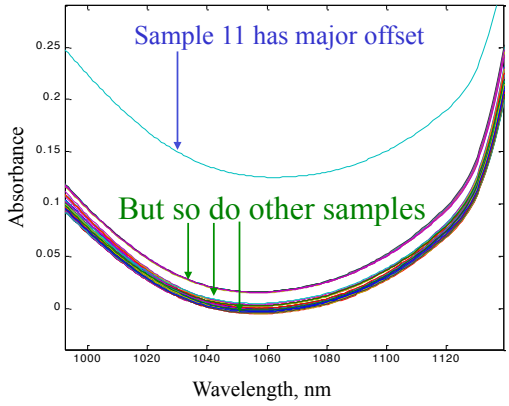
Baseline Ramp?



246

EIGENVECTOR RESEARCH INCORPORATED

Yes

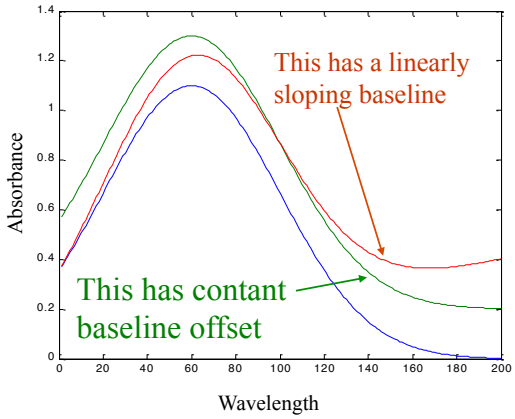


247

EIGENVECTOR RESEARCH INCORPORATED

Baseline Problems Example

Here are three identical spectra, except:

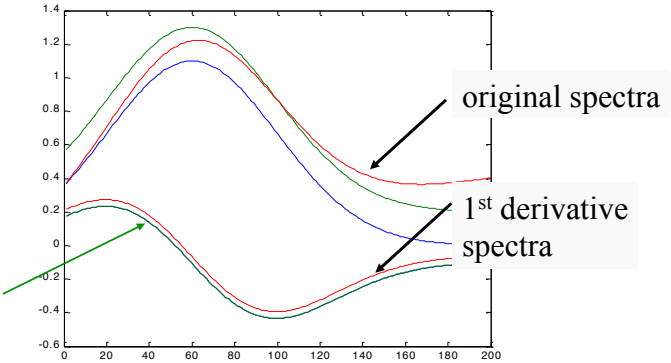


248

EIGENVECTOR RESEARCH INCORPORATED

Using Derivative Spectra

Take 1st derivative of the Three Spectra

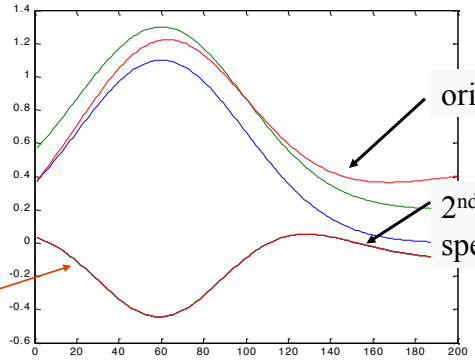


249

EIGENVECTOR RESEARCH INCORPORATED

Using Derivative Spectra

Take 2nd derivative of the Three Spectra

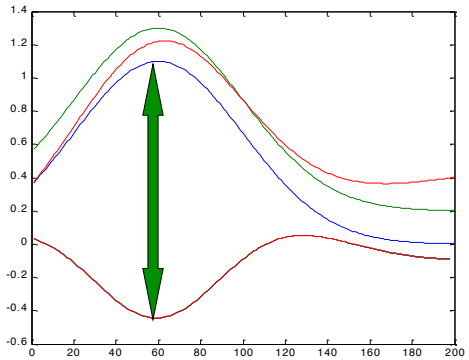


250

EIGENVECTOR RESEARCH INCORPORATED

Comments on 2nd Derivative

2nd derivative spectra look different

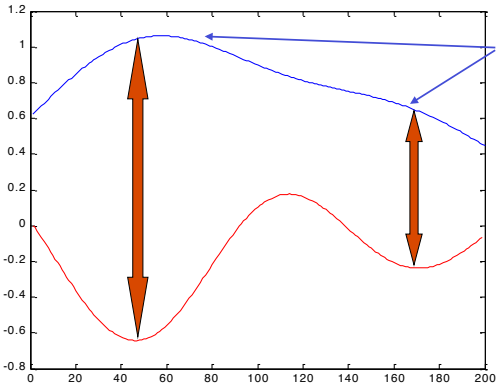


251

EIGENVECTOR RESEARCH INCORPORATED

Comments on 2nd Derivative

2nd derivative increase resolution



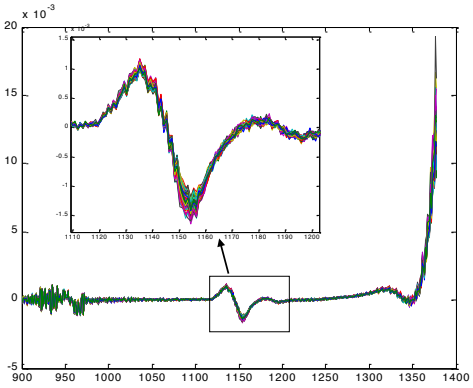
Peak with possible shoulder

2nd derivative resolves the peak and shoulder

252

EIGENVECTOR RESEARCH INCORPORATED

2nd Derivative of Learning Set Spectra



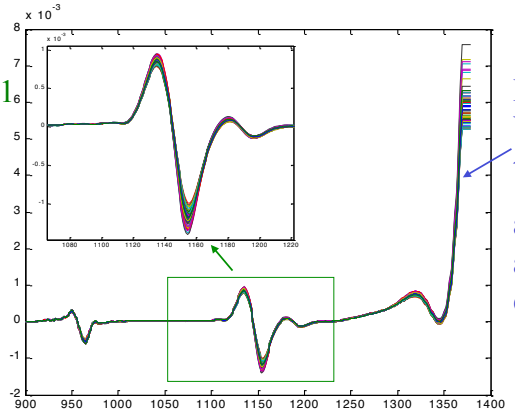
Derivatives can also increase the effect of Noise!

253

EIGENVECTOR RESEARCH INCORPORATED

Filter the Data Before Taking 2nd Derivative: Savitzky-Golay

Even Sample 11 looks good!



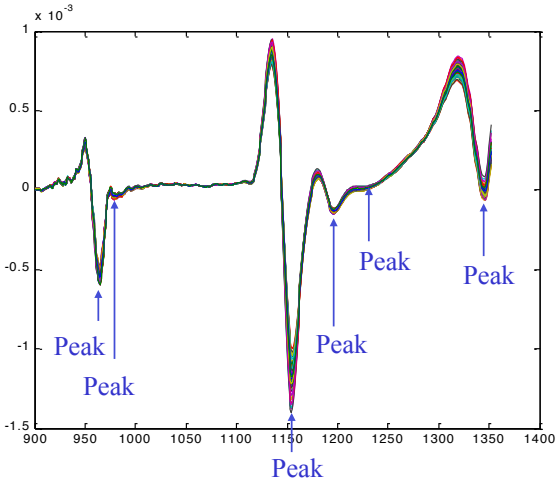
Must remove Variables in this steep area

and be careful about "end effects"

254

EIGENVECTOR RESEARCH INCORPORATED

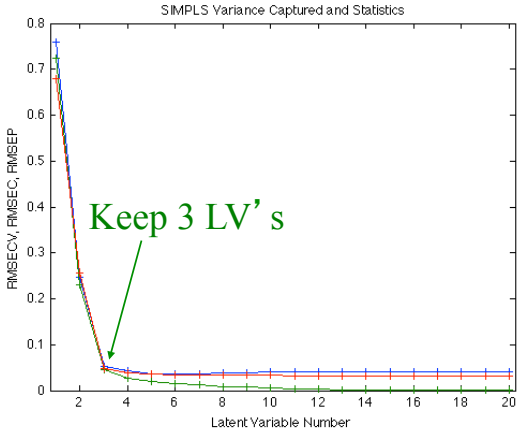
Notice the Resolution



255

EIGENVECTOR RESEARCH INCORPORATED

PLS-1 of 2nd Derivative Spectra for the Learning Set

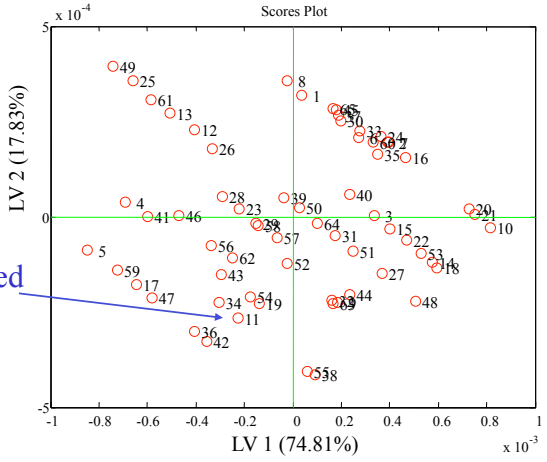


256

EIGENVECTOR RESEARCH INCORPORATED

Scores Plot

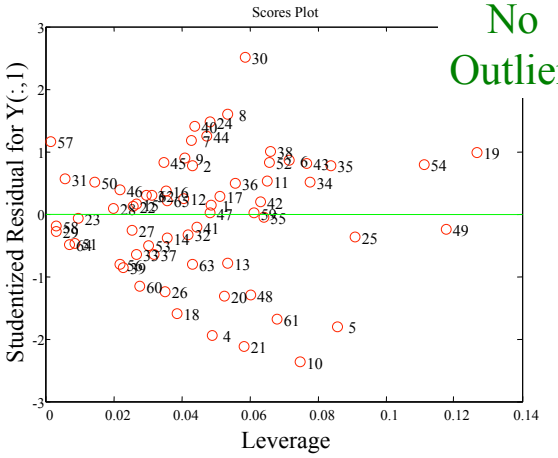
Sample 11 has returned to the fold



257

EIGENVECTOR RESEARCH INCORPORATED

Model Influence Plot

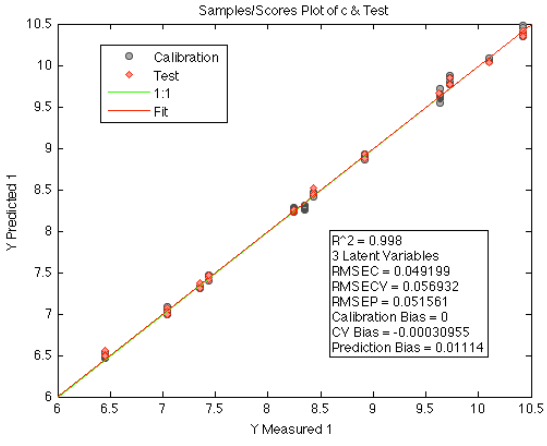


No Outliers

258

EIGENVECTOR RESEARCH INCORPORATED

Prediction of [NaOH]



259

EIGENVECTOR RESEARCH INCORPORATED

Comparison of Regression Methods and Pretreatments

	RMSECV	RMSEP	
PLS-1 2nd Derivative (3 LV)	0.0569	0.0516	Including sample 11
PLS-1 (3 LV)	0.0850	0.0825	
PLS-1 (4 LV)	0.0557	0.0708	
PCR-1 (4 PC)	0.0592	0.0743	

Need to ask: is the difference statistically significant with respect to error in the reference method.

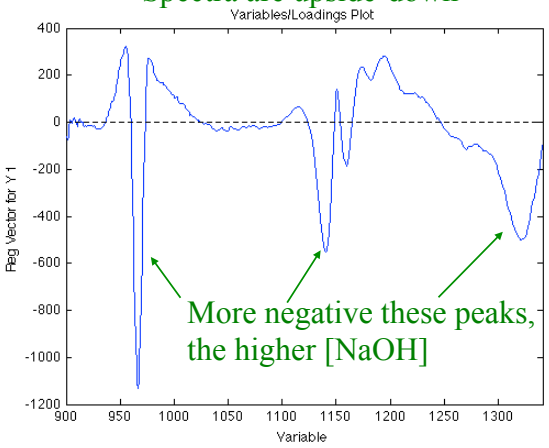
Note: It is possible to obtain a RMSEP somewhat smaller than that from the reference method.

260

EIGENVECTOR RESEARCH INCORPORATED

Regression Coefficients

Remember 2nd Derivative
Spectra are upside-down



261

EIGENVECTOR RESEARCH INCORPORATED

PLS-2

- PLS-2 is just like PLS-1 except the **Y**-Block has more than one variable (multivariate Y).
- Note that MLR and PCR can also be performed for multivariate Y, however since the **Y**-block is not used to identify the PCs it is the same as multiple univariate Y models.
- PLS uses the **Y**-block, therefore PLS-2 can provide different results than multiple PLS-1 (univariate Y) models.
- An important use of PLS-2 is PLS-DA, a classification technique

EIGENVECTOR RESEARCH INCORPORATED

We Wish to Identify Geographic Origin of Romano-British Pottery

We have samples of pottery from:

Gloucester	22
Wales	16
New Forest	10

X-Block: 8 metal concentrations as determined by Atomic Absorption Spectroscopy:

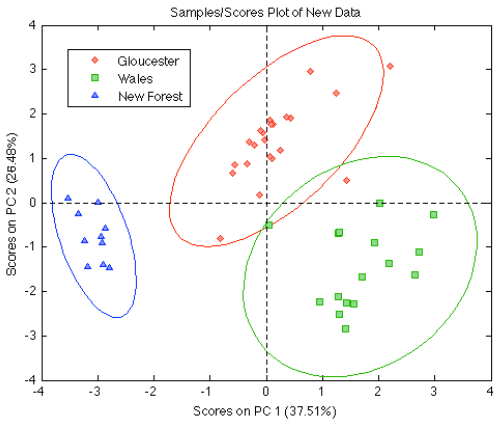
Al Ba Ca Fe K Mg Na Ti

A. Tubb, et. al., *Archacometry*, **22**, 153 (1980)

EIGENVECTOR RESEARCH INCORPORATED

PCA (autoscaled) Scores Plot for **X**-Block of All 48 Samples

Not very good discrimination.



264

EIGENVECTOR RESEARCH INCORPORATED

PLS-DA
Y-Block coding of Class Membership

Sample	Gloucester	Wales	New Forest
1	1	0	0
2	0	1	0
3	0	0	1

Not Glouchester

Not Wales

Yes New Forest

Create and assign classes to samples and software will code samples

265

EIGENVECTOR RESEARCH INCORPORATED

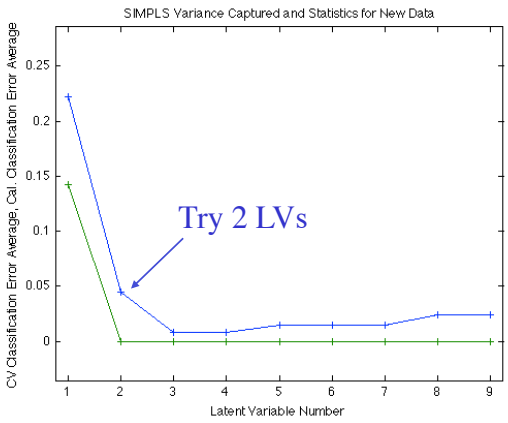
Split Samples into a Learning Set (32 samples) and Test Set (16 samples)

Learning Set:	Gloucester	15
	Wales	11
	New Forest	10
Test Set:	Gloucester	7
	Wales	5
	New Forest	3

266

EIGENVECTOR RESEARCH INCORPORATED

RMSECV (5 Split Venetian Blind) vs. Number of Latent Variables

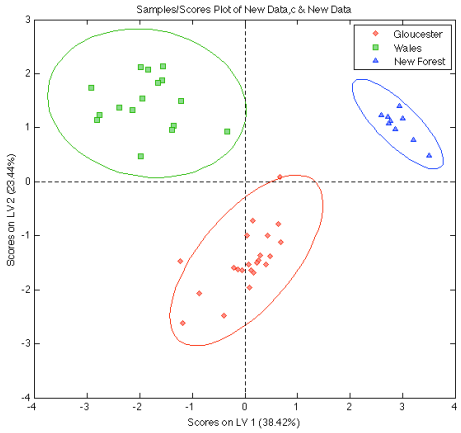


267

EIGENVECTOR RESEARCH INCORPORATED

PLS-DA Scores PV2 vs LV1

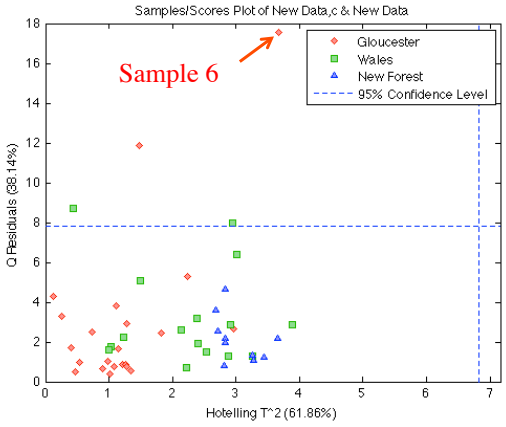
Bringing in the Y-block during the creation of the LVs improved class discrimination



268

EIGENVECTOR RESEARCH INCORPORATED

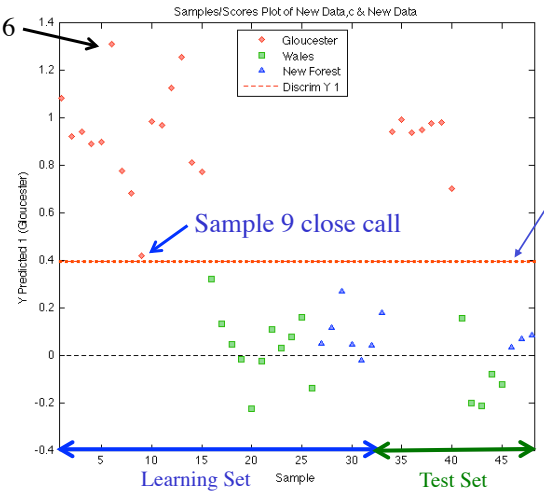
Always Check X-block Influence Plot



269

EIGENVECTOR RESEARCH INCORPORATED

Predictions for Class Gloucester



Sample 6

No false positives or negatives

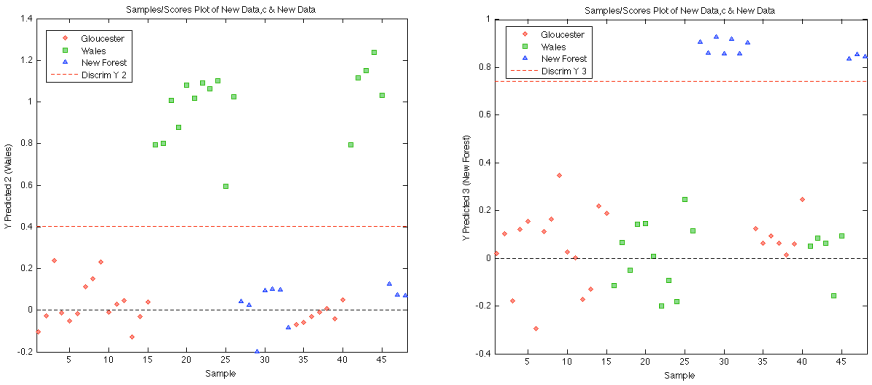
Sample 9 close call

Decision Line

270

EIGENVECTOR RESEARCH INCORPORATED

Predictions for Class Wales

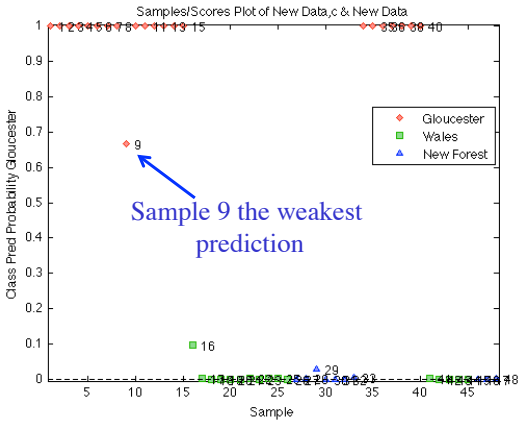


No false positives or negatives

271

EIGENVECTOR RESEARCH INCORPORATED

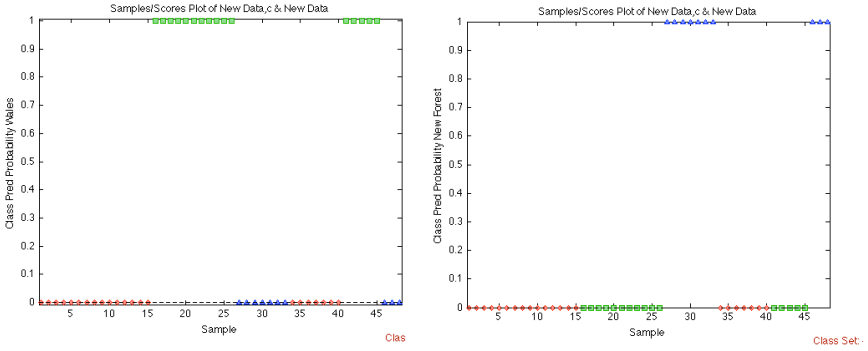
Class Prediction Probability for Gloucester



272

EIGENVECTOR RESEARCH INCORPORATED

Class Prediction Probability for Wales and New Forest



273

EIGENVECTOR RESEARCH INCORPORATED

Dangers in using PLS-DA

Create a Learning Set X-Block of 100 samples with 100 random variables

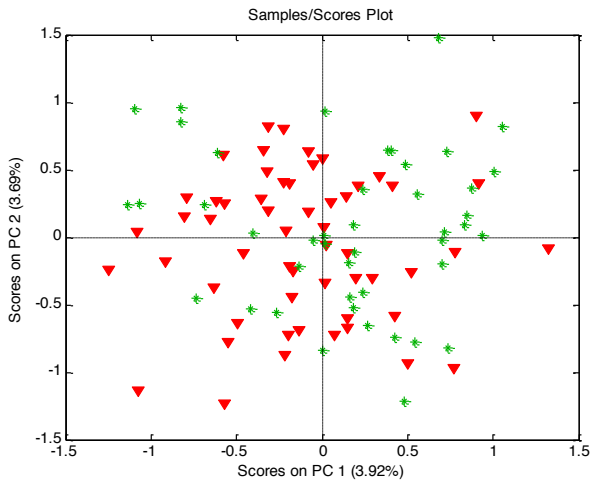
Assign samples to two classes at random – Y-Block

Create a similar Test Set of 100 samples

274

EIGENVECTOR RESEARCH INCORPORATED

PCA Scores Plot of Learning Set

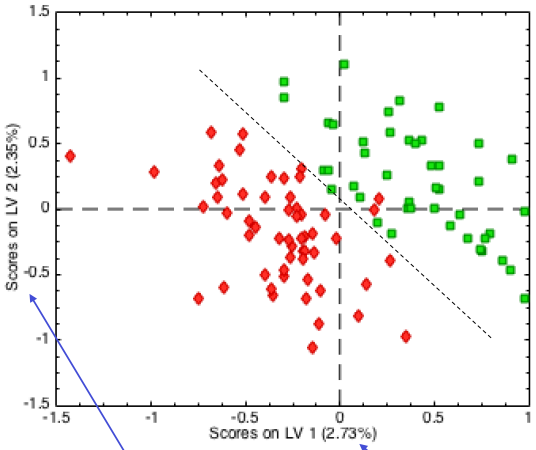


Totally Random distribution of samples and classes

275

EIGENVECTOR RESEARCH INCORPORATED

PLS-DA Scores Plot of Learning Set



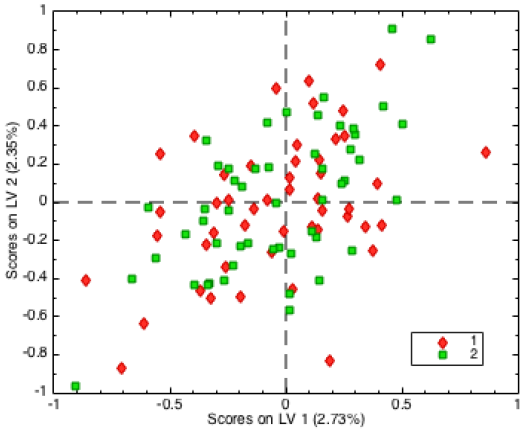
PLSDA has found differences between the two classes in the Learning Set

Notice small amount of variance captured.

276

EIGENVECTOR RESEARCH INCORPORATED

The Test Set Projected onto the Model Scores Space



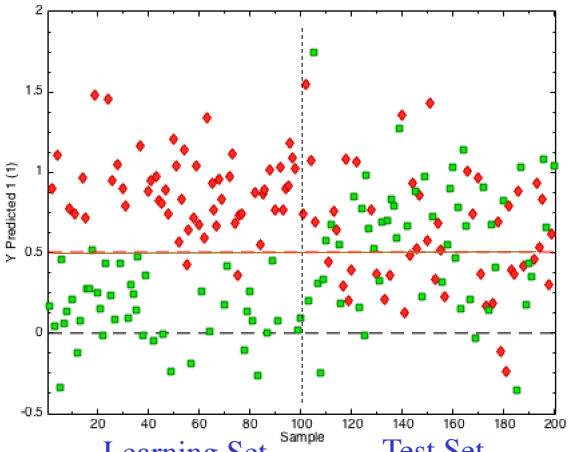
The PLSDA Model fails to distinguish the samples in the Test Set

This is important. Validation can indicate performance problems – do not kid yourself!

277

EIGENVECTOR RESEARCH INCORPORATED

Classification of Learning Set and Test Set



Learning Set Test Set

278

EIGENVECTOR RESEARCH INCORPORATED

Always Use an Independent Test Set to Validate Your Model(s)!!!

279

EIGENVECTOR RESEARCH INCORPORATED

Even the Most Expensive Cars Require Regular Maintenance

- Periodically Revalidate Model to make sure conditions have not changed over time
- Update Model if necessary
  - many new tools to assist in this process.

280

EIGENVECTOR RESEARCH INCORPORATED

Also Remember

Rome Was Not Built in a Day

It took days, months or even years to take the data.

It may take days or weeks to build a proper model. Initial models may indicate that new measurements are necessary.

Be Patient and Think about What You Are Doing

Math, Physics and Chemistry are your guides.

281

LVC

EIGENVECTOR RESEARCH INCORPORATED