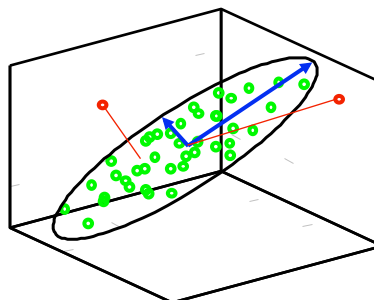


Chemometrics

Without Equations
(or hardly any)



©Copyright 2004, 2015
Donald B. Dahlberg and Eigenvector Research, Inc.
No part of this material may be photocopied or reproduced in
any form without prior written consent from Eigenvector
Research, Inc. or Donald B. Dahlberg



Contact Information

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

Barry M. Wise
Eigenvector Research, Inc.
196 Hyacinth Road
Manson, WA 98831
phone: (509) 662-9213
fax: (509) 662-9214
bmw@eigenvector.com



Biometrics

Technometrics

Econometrics

Psychometrics

Chemometrics

↑
Chemical

↑
Measurements

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

3



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



This Workshop Concentrates on Two Aspects of Data Analysis:

- Exploratory Data Analysis and Pattern Recognition
- Regression

5



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



Outline (Part 1)

- [Introduction](#)
- Exploratory Data Analysis & Pattern Recognition Motivation
- Principal Components Analysis
- SIMCA
- Summary

7



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



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

9



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.

10



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?

11



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**?

12



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



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!

NIR Hyperspectral image

256x256 pixels = 65536 samples from
126 channels / wavelengths

8,257,536 Pieces of Information!

14



Most of This Information is Either:

- **Irrelevant** to the problem we want to solve
- **Redundant**
 - problem for some types of analysis
 - 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

15



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



16



Perhaps Count the Number of Feet?



Fluffy	4
Spot	4
Puff	4

These numbers are all the same.

No Variance in this Variable



Rex	4
Lassie	4
Rover	4

Therefore no information in these numbers.

17



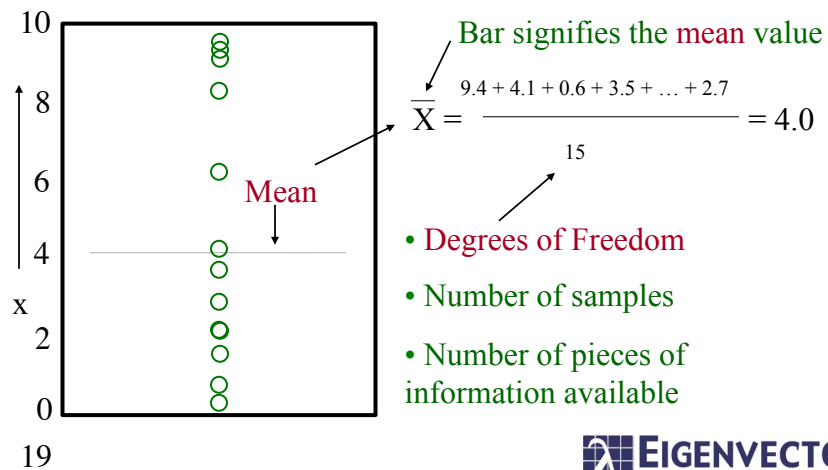
A number only has value when it is compared with other numbers

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

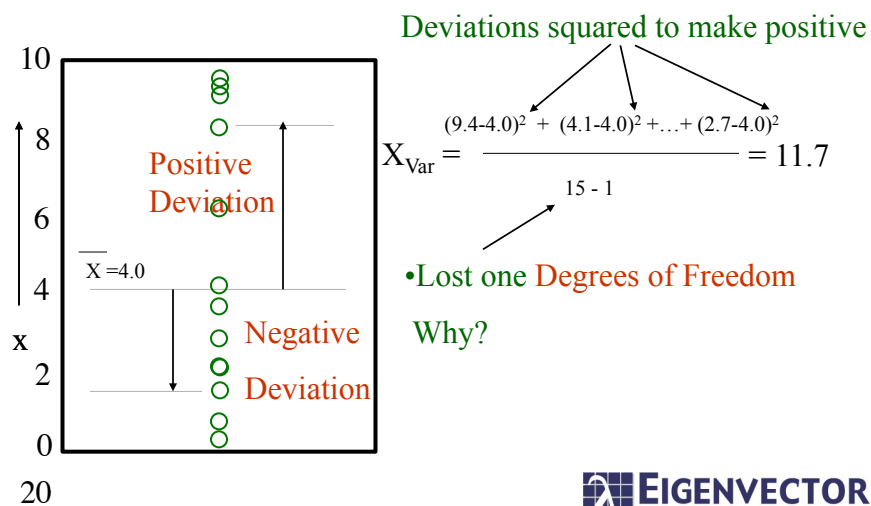
18



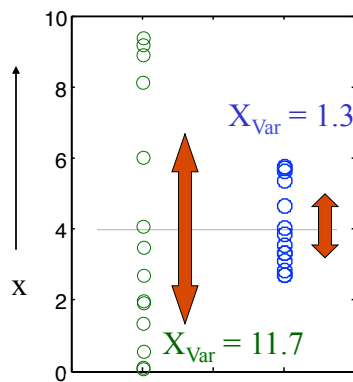
Before Defining Variance Need to Define Mean:



Calculating Variance



Both data sets have the same **mean**



but they have different **variance**

Note: Standard Deviation

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

21



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)

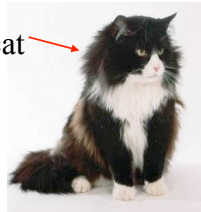
Back to the Cats and Dog problem

22



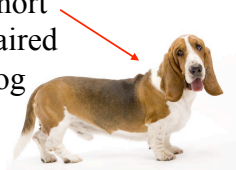
Perhaps length of fur can be used to tell cats from dogs?

Long haired cat



Large Variance

Short haired dog



Long haired dog



Short hair cat



23

But still little
Discrimination
Power between cats
and dogs.



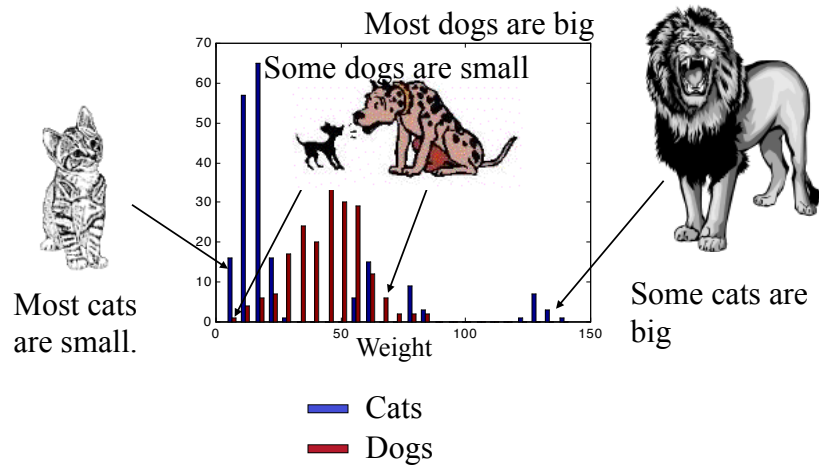
Variance = Information

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

24



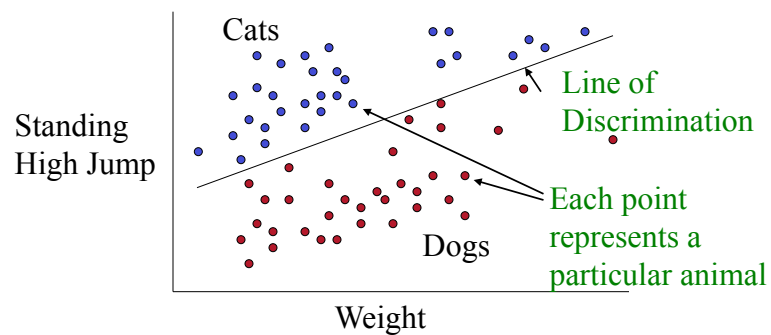
Perhaps Weight of Animal?



25



Perhaps two variables *Together*?



26



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...

27



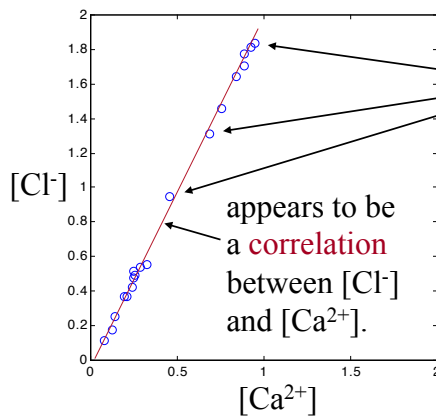
Table of 20 Solutions Analyzed for Ca^{2+} and Cl^- Concentrations.

		[Ca^{2+}]	[Cl^-]	20 x 2 Matrix table of numbers
20 Rows Samples	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	
			2 Columns Variables	

28



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



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.

29



How to find this underlying variable?

Next topic:

Principal Component Analysis

PCA

30



Outline (Part 1)

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

©Copyright 2004, 2015
Donald B. Dahlberg and Eigenvector Research, Inc.
No part of this material may be photocopied or reproduced in
any form without prior written consent from Eigenvector
Research, Inc. or Donald B. Dahlberg

31



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

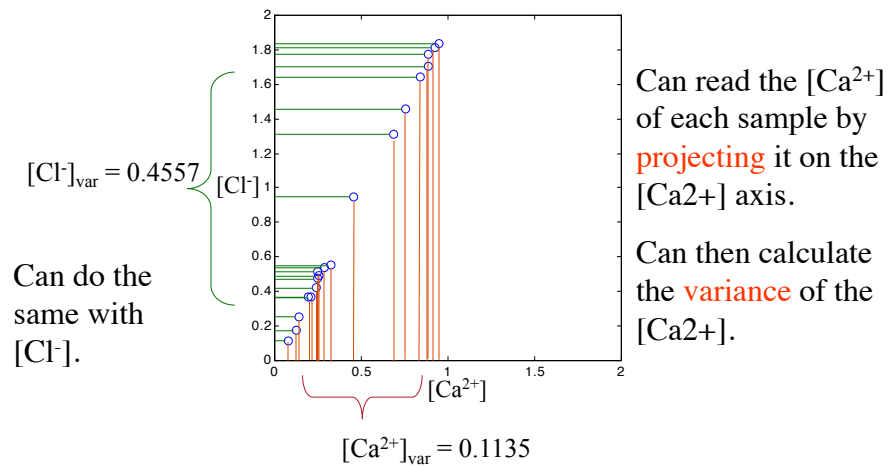
Is there a way to concentrate the
information from *many* variables into a *few*
(hopefully 2 or 3) underlying variables?

Hint: Variance = Information

32



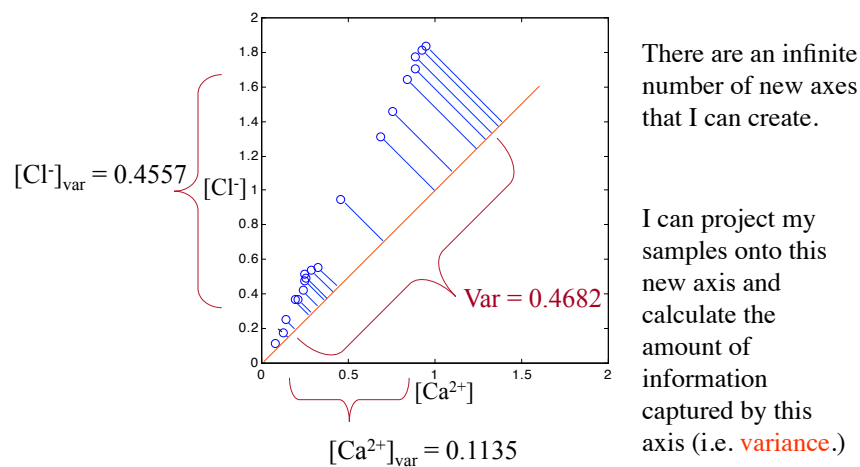
Plot the Cl^- Concentration vs. the Ca^{2+} Concentration



33



Find a New Axis of Maximum Variance:

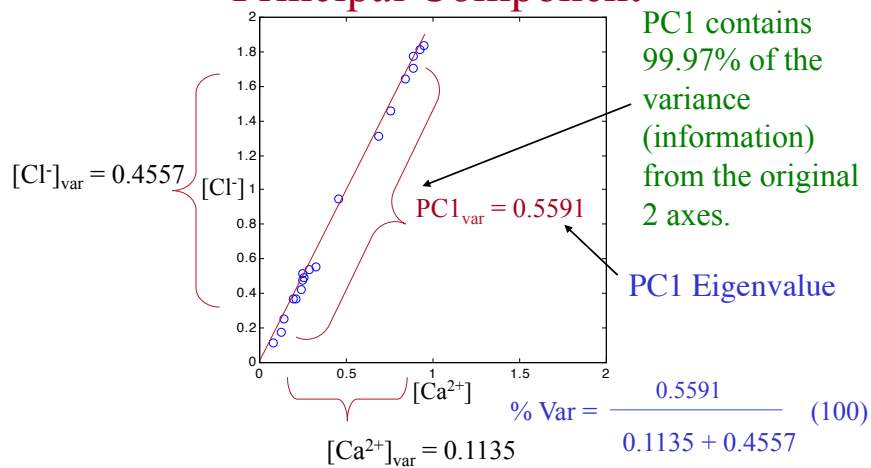


34



Only One New Axis Which Captures the
Maximum Variance:

Principal Component

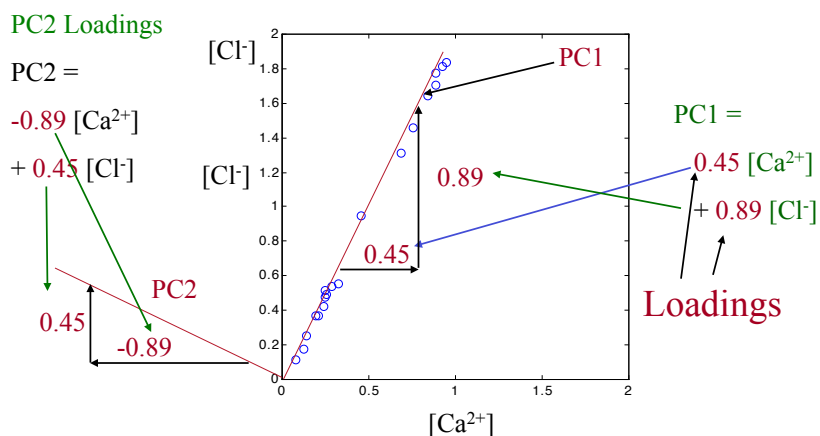


35



How Was PC1 Constructed?

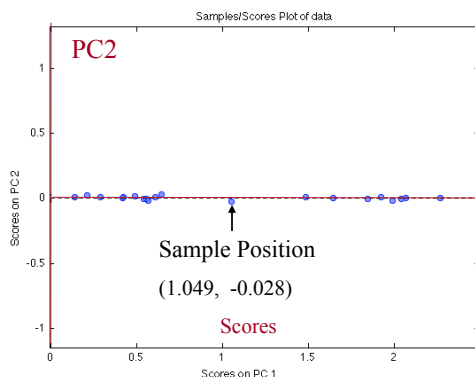
Linear Combination of Original Axes



36



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)

37



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

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

- **Loadings** for PC1 in a ratio of 1:2
 - One part Ca^{2+} and two parts Cl^-
- PC1 related to $[\text{CaCl}_2]$
 - $\sim 99.97\%$ variance is systematic
- PC2 related to random measurement error
 - $\sim 0.03\%$ noise

38



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: $[\text{Ca}^{+2}]$ and $[\text{Cl}^-]$
 - 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.

39



Summary of What We Have Done and Learned

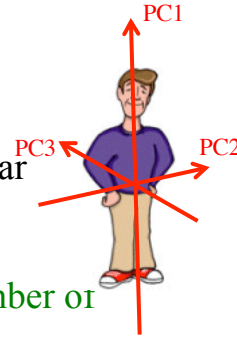
- **Variance** = Information
- Many variables contain information, but much of the information may be redundant (**correlated**) or irrelevant.
- Use linear combinations of the original variables to create new variables (**Principal Components**) that combines redundant information → **PCA**
- The **Rank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

40



How Does PCA Find the PC's?

- 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**.
 - $PCs \leq \min(\text{number of samples, number of variables}) = \text{mathematical rank of the data}$



41



What Does PCA Give Me?

- Most of the **variance** (information) is concentrated in the first few PC's.
 - Some may be relevant to the problem of interest
- **Small random noise** is sifted into the later PC's
 - 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

42



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



Say 100 Times Before You Go to
Sleep Tonight!!

Loadings ↔ Variables

Scores ↔ Samples

Eigenvalues ↔ Variance

44



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.

<u>Quarry</u>	<u>Number of Samples</u>	<u>Row Index</u>
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).

45



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
•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•
•	•	•	•	•	•	•	•	•	•	•
20	880	156	36	279	388	37	103	15	53	143

20 x 10 Matrix

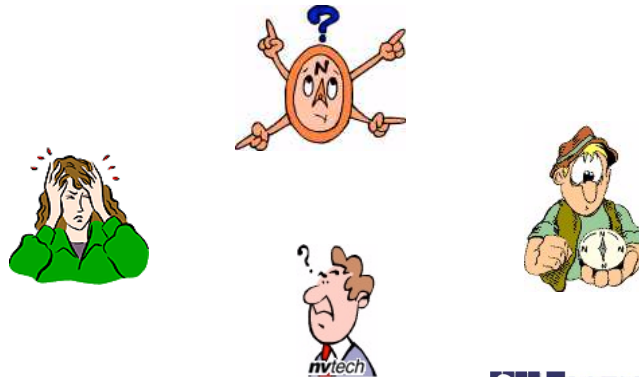
*X-Ray Fluorescence (ppm)

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

46



Plot Each Sample as a Point in 10-Dimensional Space

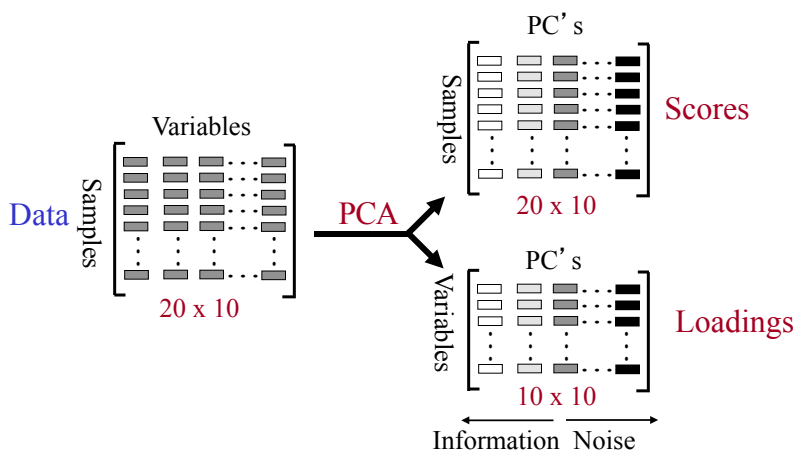


47

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



Before PCA do:
Data Preprocessing

- None
- Mean centering
- Autoscaling

50



PCA of GLASS - No Preprocessing

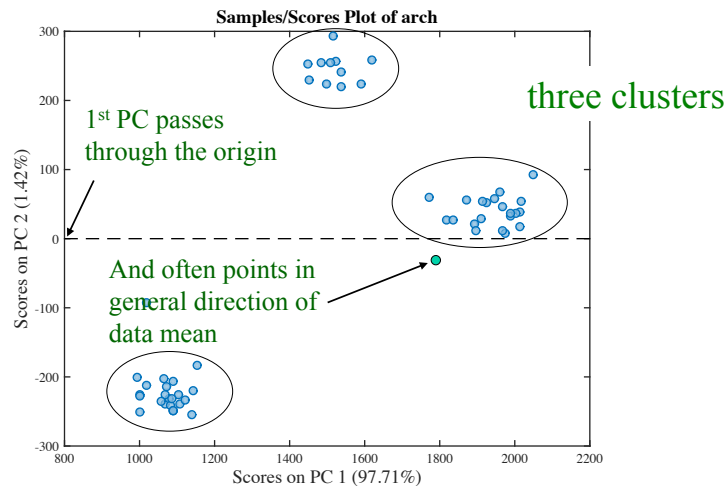
Principal Components Analysis Model
X-block: arch 56 by 10
Included: [1-56] [1-10]
Preprocessing: None
Num. PCs: 5

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

51



PCA Scores Plot - No Preprocessing

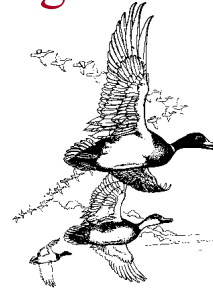


52



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!!

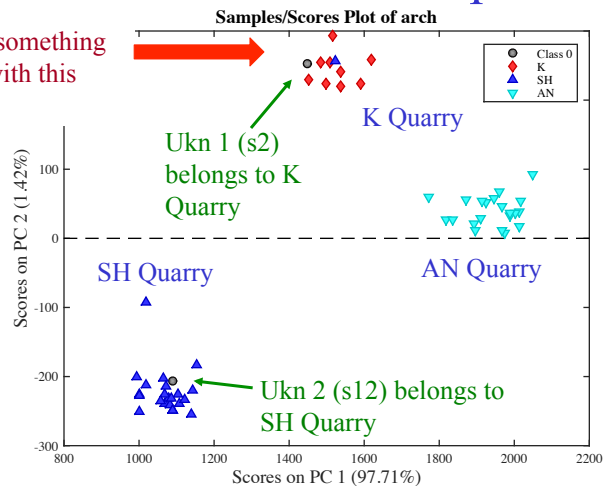


53



PCA Scores Plot - No Preprocessing

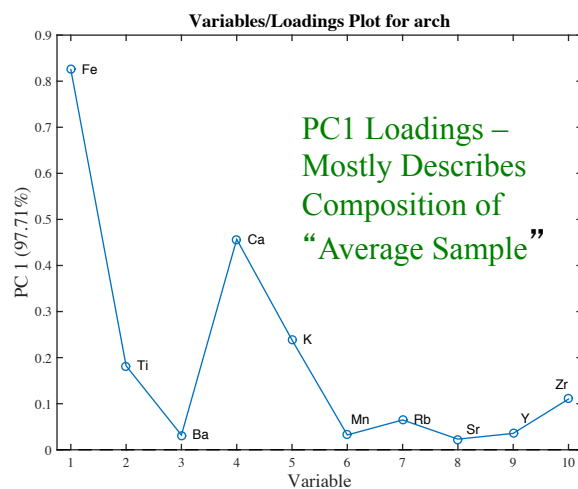
Is there something wrong with this picture?



54



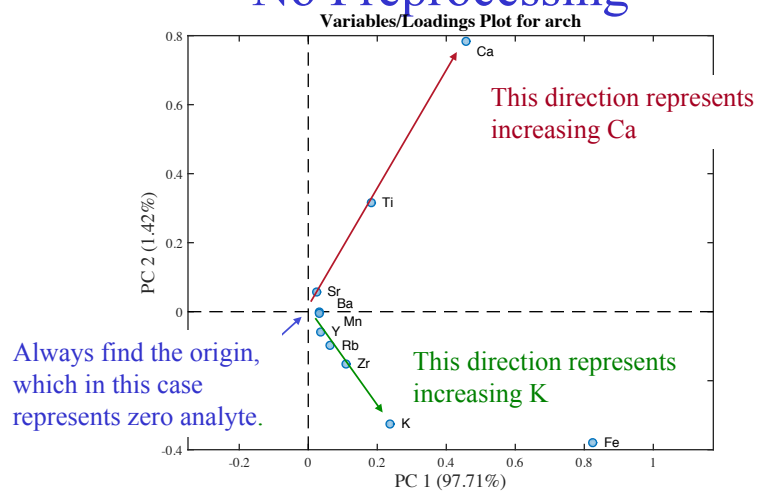
Loadings Plot



55



PCA Loading Plot PC2 vs PC1 - No Preprocessing



56



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



57

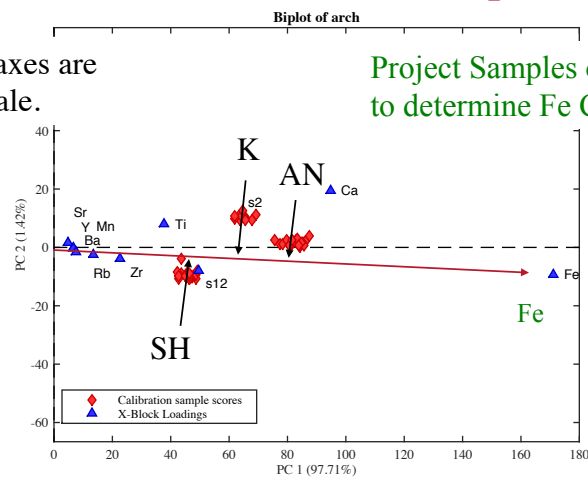


Easier to Understand by Superimposing
Scores and Loadings - Biplot

Note both axes are
equal in scale.

Project Samples onto Fe line
to determine Fe Content.

PC 1+2
~99.1%



58



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.

59



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...

60



Data Preprocessing

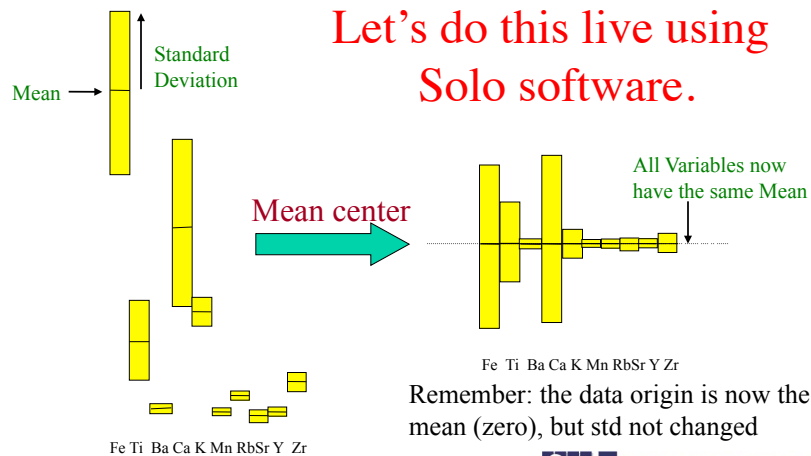
- None
- Mean centering
- Autoscaling

61



Repeat the PCA with Mean Centered Data

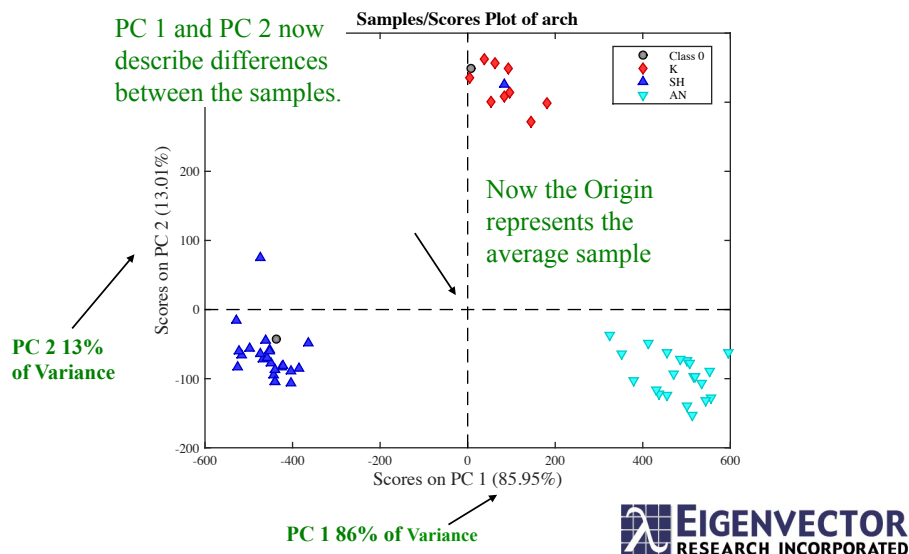
Let's do this live using Solo software.



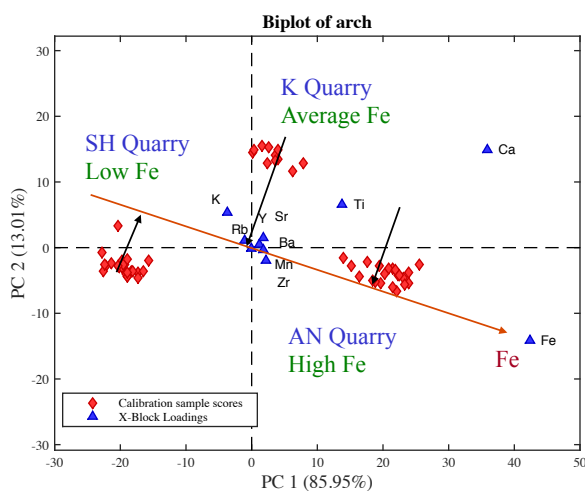
62



Scores Plot - Mean Centering



Biplot - Mean Centering



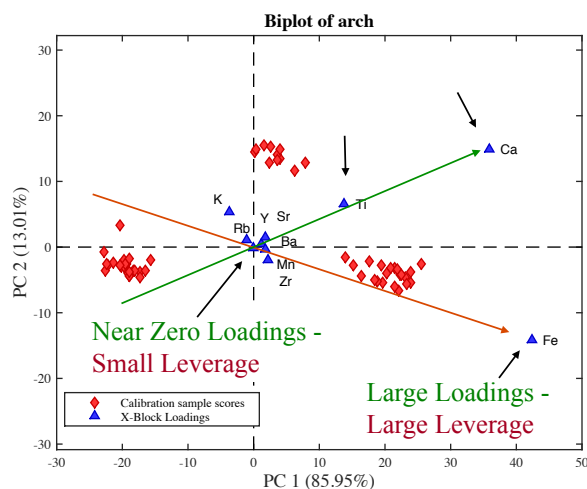
Note axes have equal scale

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.

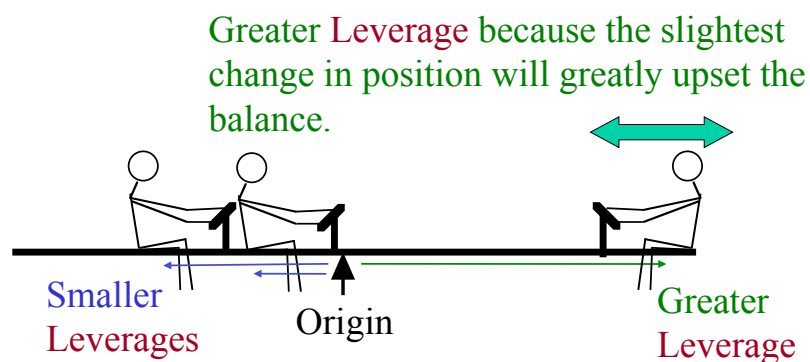
Large Leverage

65



[Return to Slides](#)

Leverage: Distance from Origin



Leverage is measured by a quantity T^2 .

66



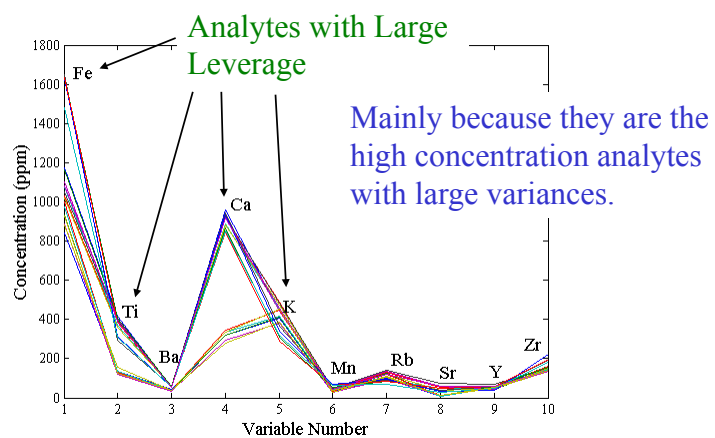
Variables with Small Leverages:

- 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



Plot the Data: Analyte Profile



68



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



69

Much More
Fair!



 **EIGENVECTOR**
RESEARCH INCORPORATED

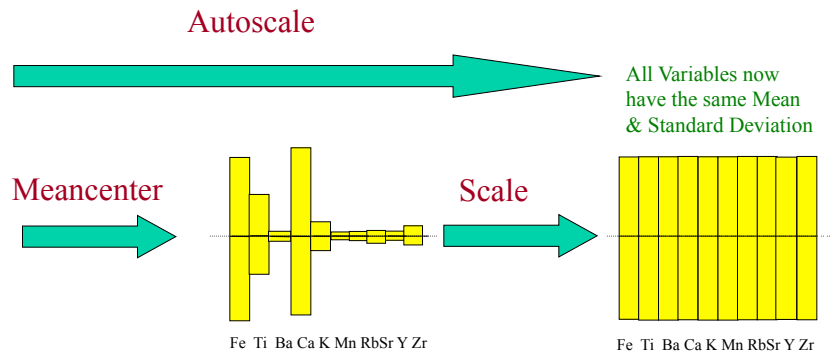
Data Preprocessing

- None
- Mean centering
- Autoscaling

70

 **EIGENVECTOR**
RESEARCH INCORPORATED

Autoscaling



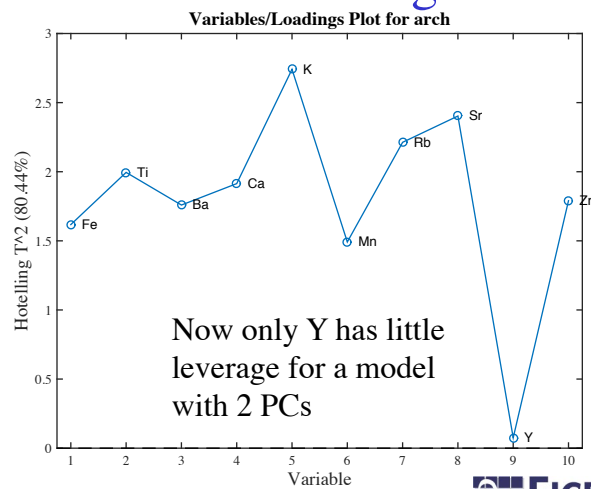
Remember: autoscaling includes mean centering

Back to Solo



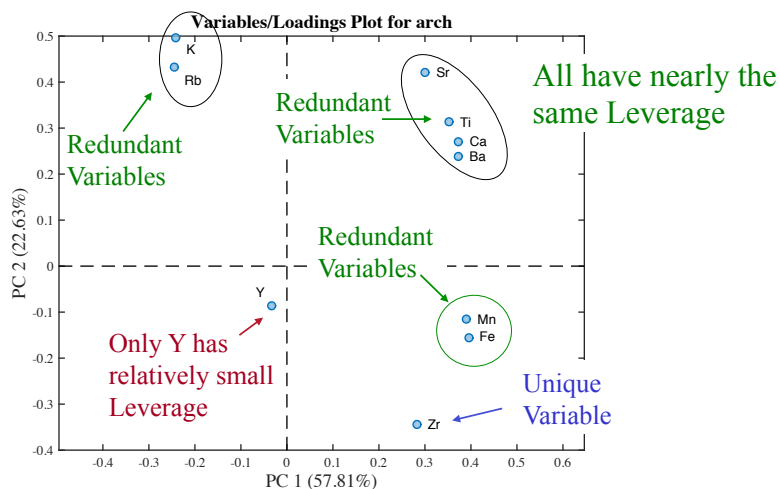
71

Leverages of Variable after Autoscaling



72

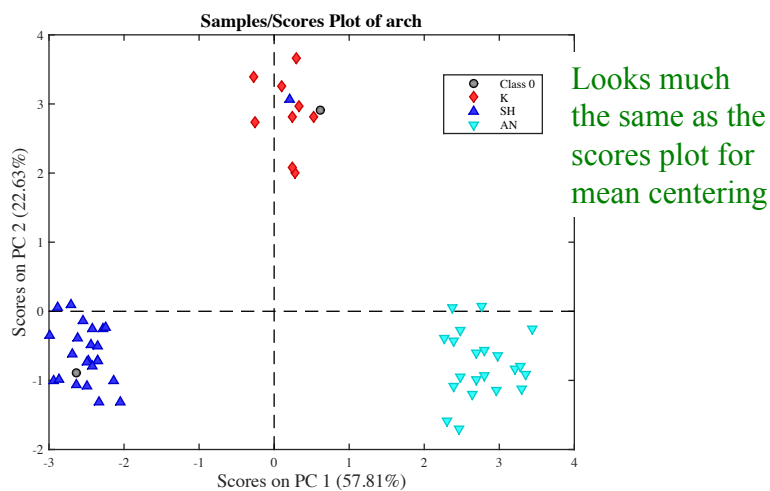
Loadings Plot - Autoscaling



73



PCA Scores Plot -Autoscaling

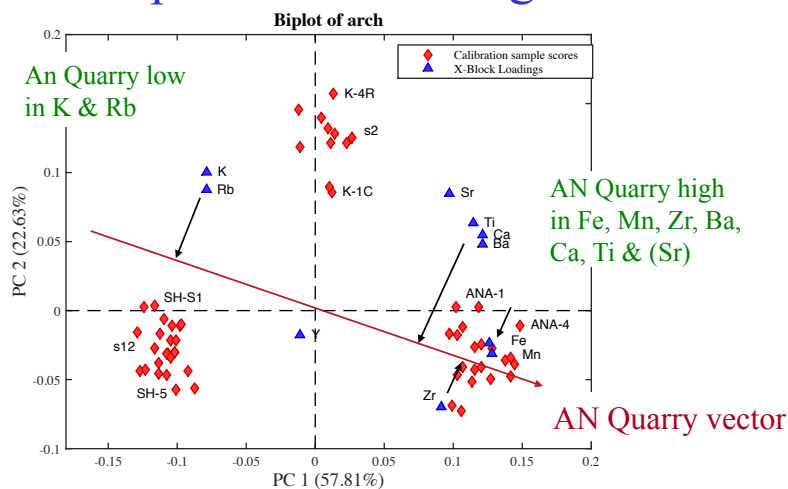


74



Return from Slides

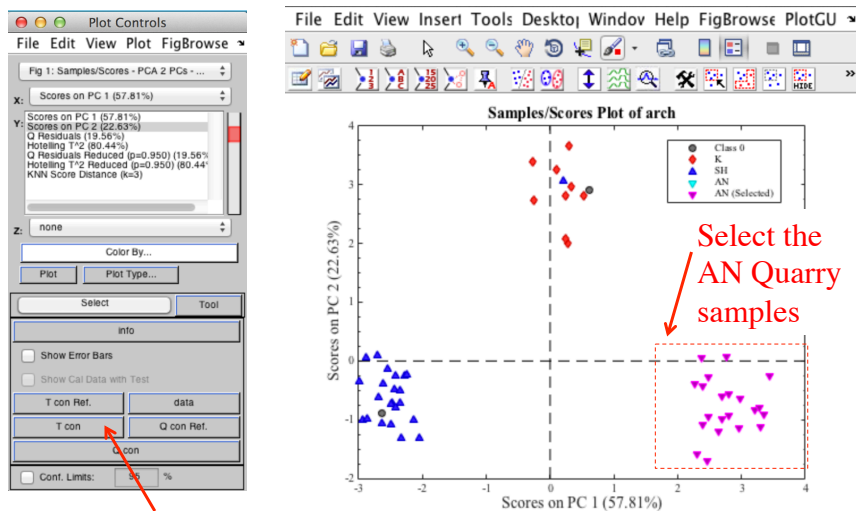
Biplot - Autoscaling



75 Back to Solo to see an easier way



T- Contribution of Variables to Samples

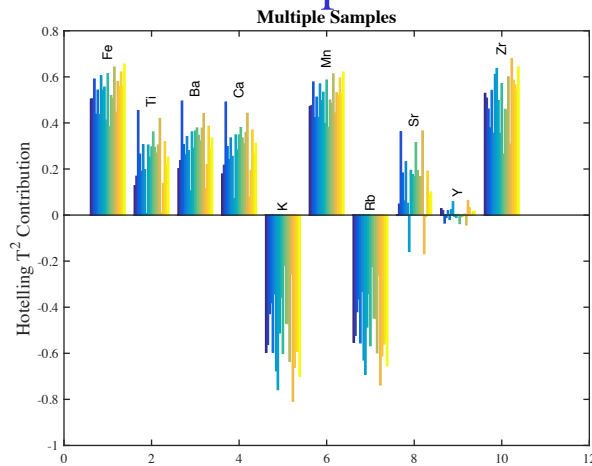


Choose T con

76



Variable Leverage of C Quarry Samples



77



[Return to Slides](#)

What Can I See From Biplot or T-Con?

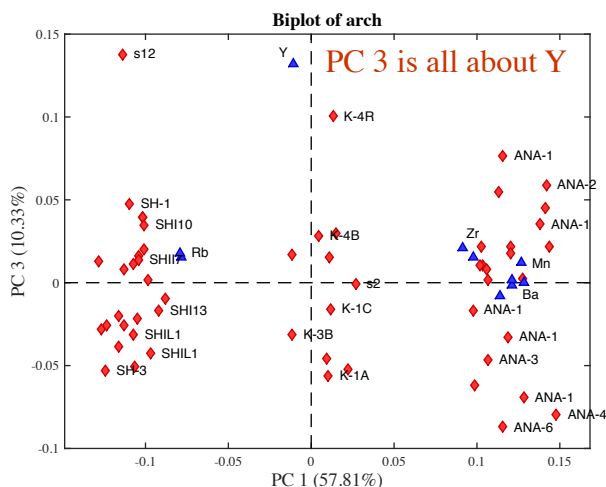
(In 2PC - Autoscaled space)

- 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 and Sr
 - Moderately high in Ti, Ba & Ca
 - Low in Zr.

78



Also Check Higher PCs



PC 3 Scores
show no
discrimination
of samples by
quarry

79



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

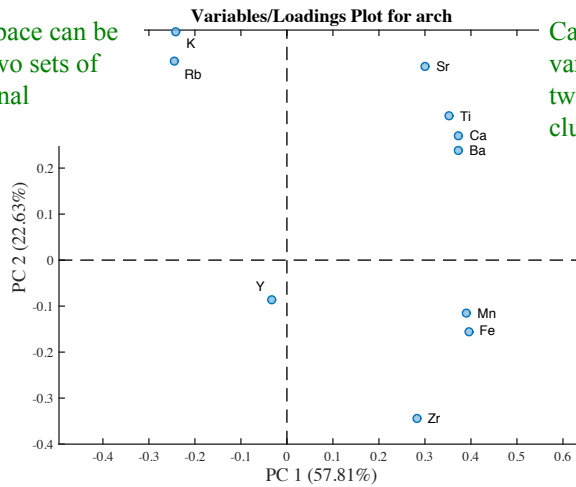
80

Go to Solo



Examine Loadings Plot Again

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

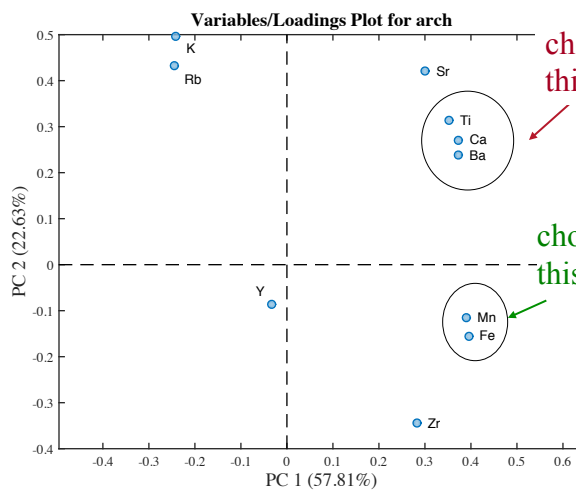


Can choose one variable from each of two ~ orthogonal clusters.

81



Examine Loadings Plot Again



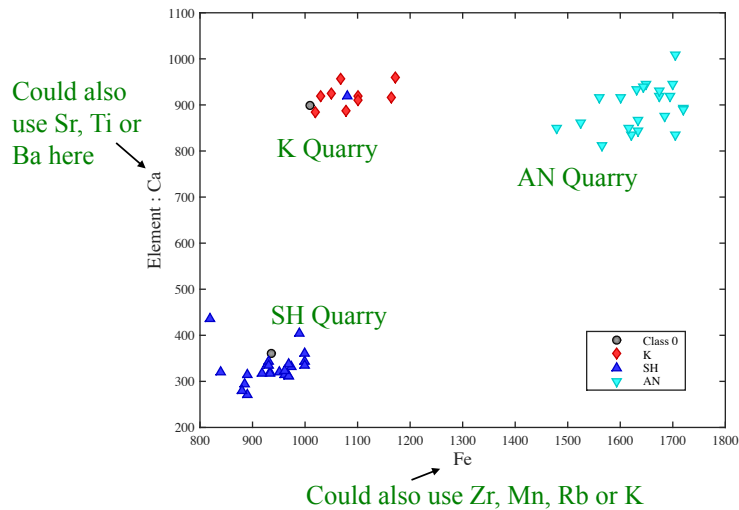
choose Ca from this group

choose Fe from this group

82



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



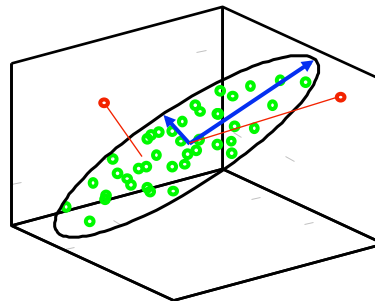
83



Return from Solo

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



Yet a Little More Complicated Problem

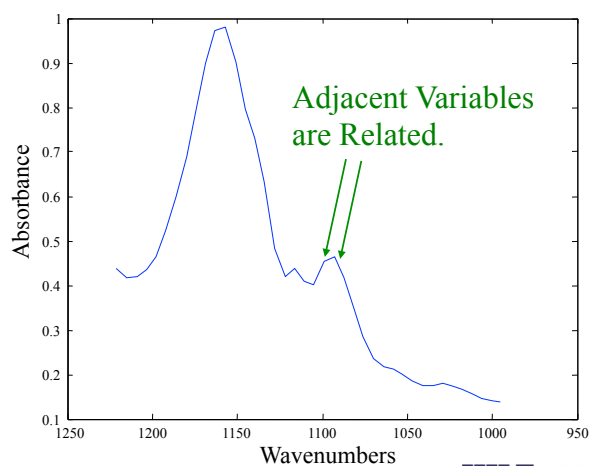
- We wish to use FT-IR spectra and pattern recognition to distinguish authentic olive oil from counterfeit or adulterated olive oil.
- This time we shall see some special properties associated with **Spectral Data**.
- We shall also learn a new pattern recognition technique called **SIMCA**.

©Copyright 2004, 2015
Donald B. Dahlberg and Eigenvector Research, Inc.
No part of this material may be photocopied or reproduced in any
form without prior written consent from Eigenvector Research,
Inc. or Donald B. Dahlberg

85



What is Special about Spectral Data?



86



- 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.

87



Olive Oil Samples Learning Set:

Corn Oil	9 samples	(#1-9)
Olive Oil	15 samples	(#10-24)
Safflower Oil	8 samples	(#25-32)
Corn Margarine	4 samples	(#33-36)

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

88



Test Set:

Corn Oil*	9 samples	(#1-9)
Olive Oil*	15 samples	(#10-24)
Safflower Oil*	8 samples	(#25-32)
Corn Margarine*	4 samples	(#33-36)
Corn Oil in Olive Oil 5, 10, 20, 30 & 40%	5 samples	(#37-41)
Almond Oil	1 sample	(#42)
Peanut Oil	1 sample	(#43)
Sesame Oil	1 sample	(#44)

* New Samples

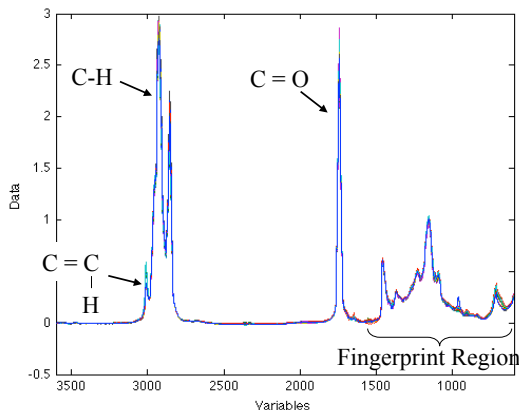
89



FTIR Spectra of 36 Sample Learning Set

Notice how spectra look alike.

Go to Solo



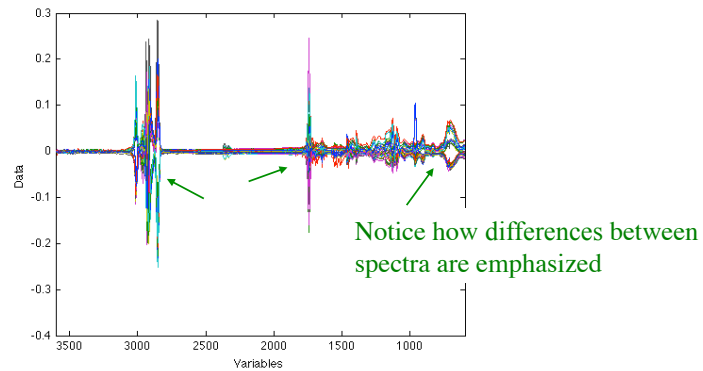
90

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



Do PCA on the Learning Set

- Examine full wavelength range
- Mean-center

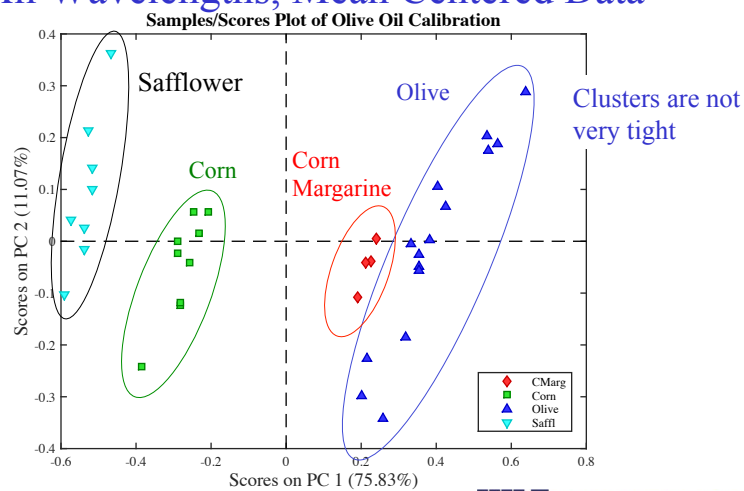


91



Scores Plot

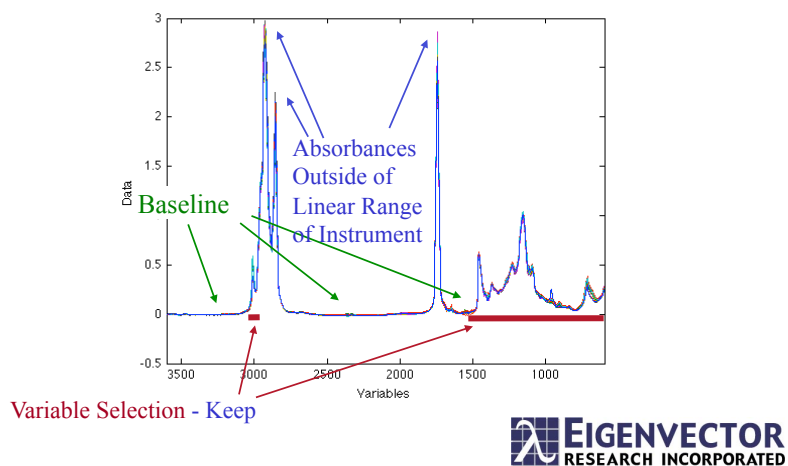
All Wavelengths, Mean Centered Data



92

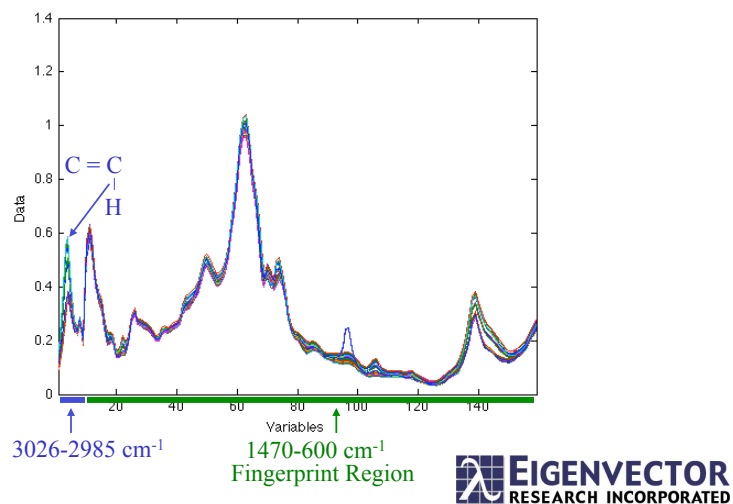


How Can the Model be Improved?



93

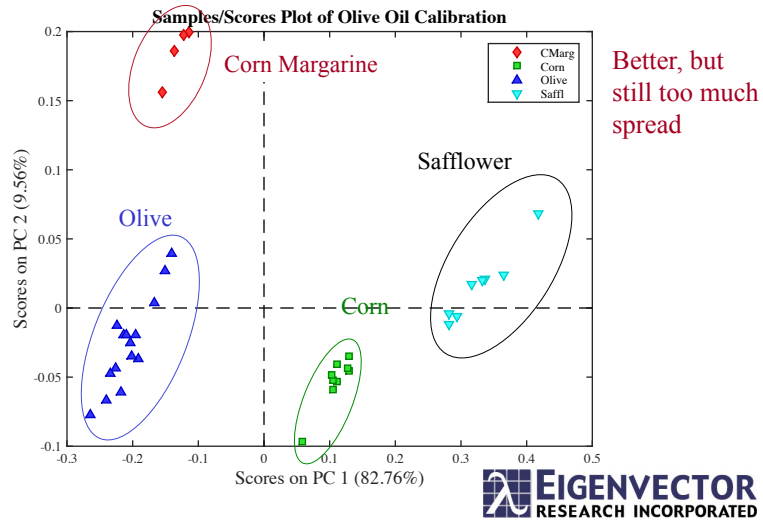
FTIR Spectra of Learning Set Selected Variables



94

Scores Plot

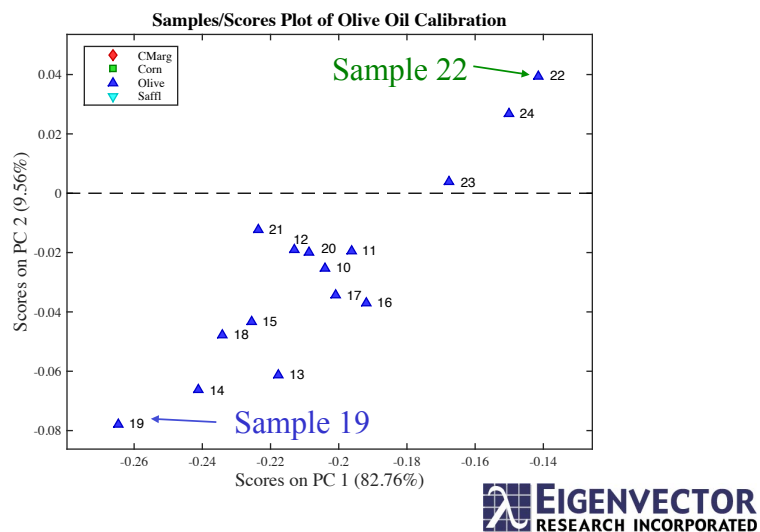
Selected Wavelengths, Mean Centered Data



95

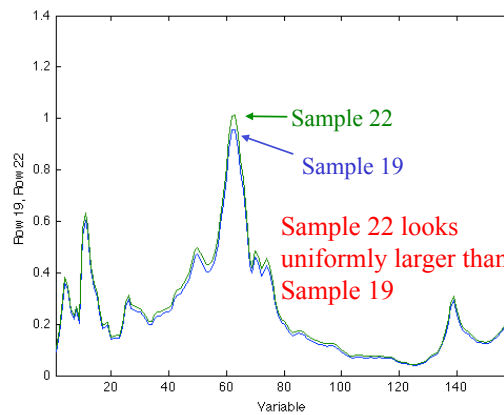
Scores Plot

Zoom In on Olive Samples



96

Spectra (Selected Wavelengths) Samples 19 & 22

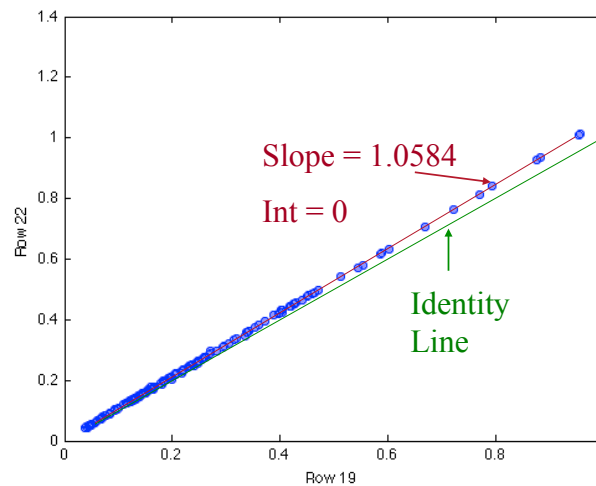


97



Return to Slides

Plot Sample 22 vs. Sample 19



98



Multiplicative Effect

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

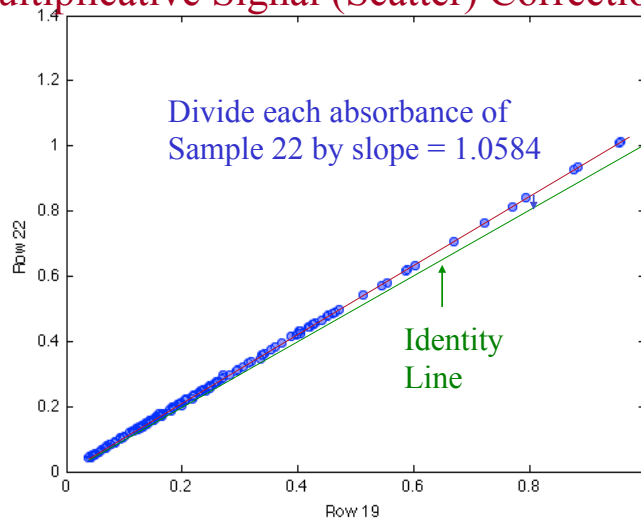
- Changing sample pathlength, *e.g.* changing light scattering with particle size.
- Changing sample density, *e.g.* changing temperature of sample.
- Changing gain of the instrument.

99



MSC

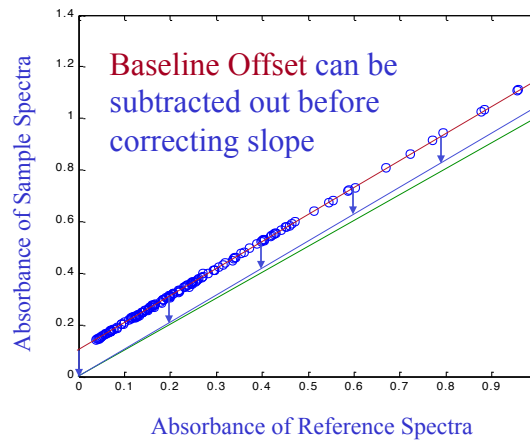
Multiplicative Signal (Scatter) Correction



100



If there is also an Offset



101



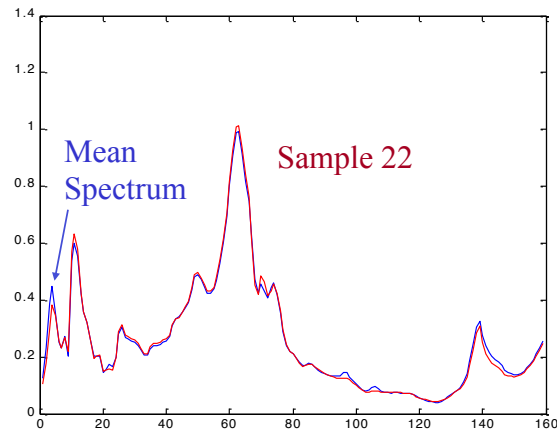
What to use as a Reference Spectrum?

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

102



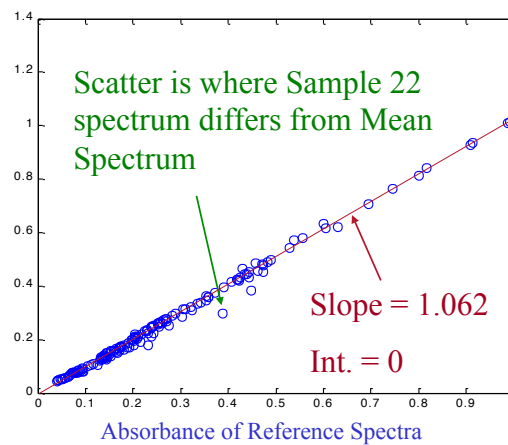
Mean Spectrum and Sample 22



103



Sample 22 vs. Mean Spectrum



104



Applying MSC

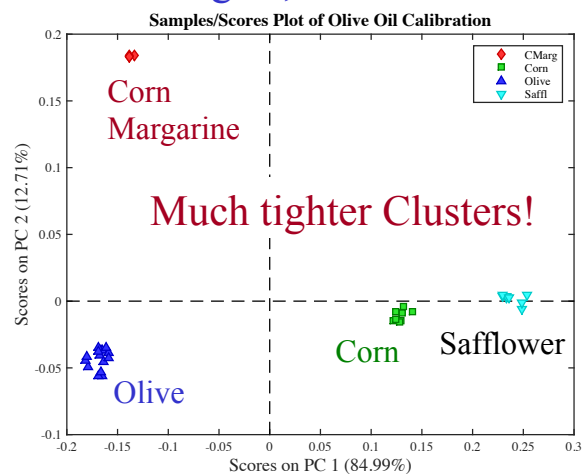
- Software will do MSC on each sample of the Learning Set
- You need to specify the Learning Set and the Reference Spectrum
 - usually mean spectrum of Learning Set

Let's try it in Solo

105



Scores Plot Selected Wavelengths, MSC & Mean Centered

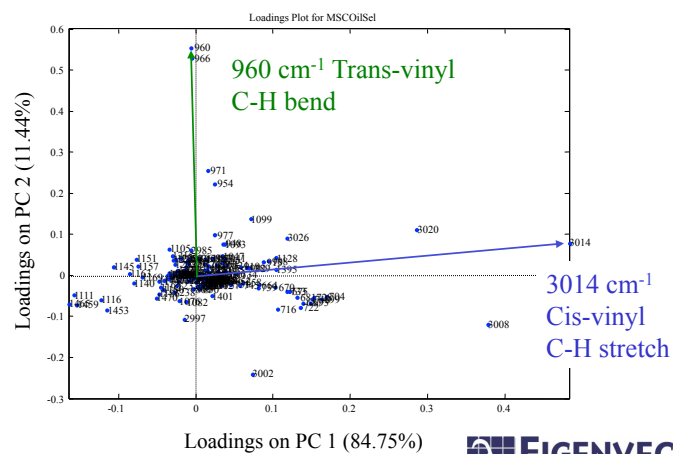


106



Loadings Plot

Selected Wavelengths, MSC & Mean Centered

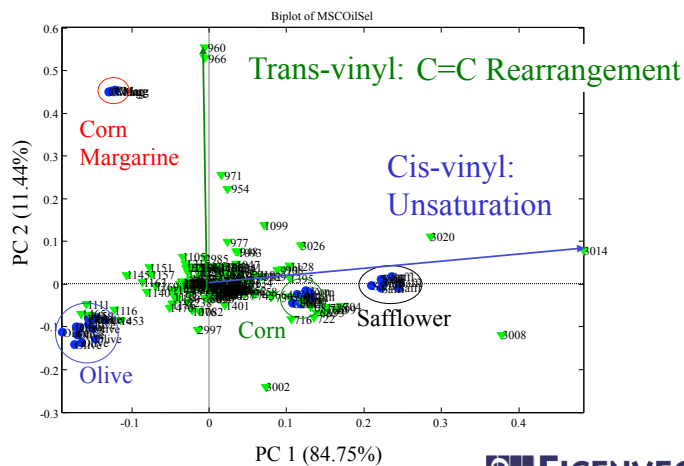


107



Biplot

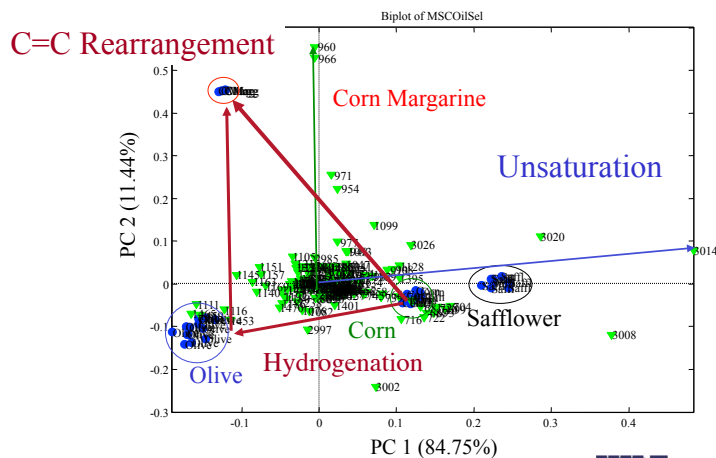
Selected Wavelengths, MSC & Mean Centered



108



Biplot Selected Wavelengths, MSC & Mean Centered



109



Return to Slides

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



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



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

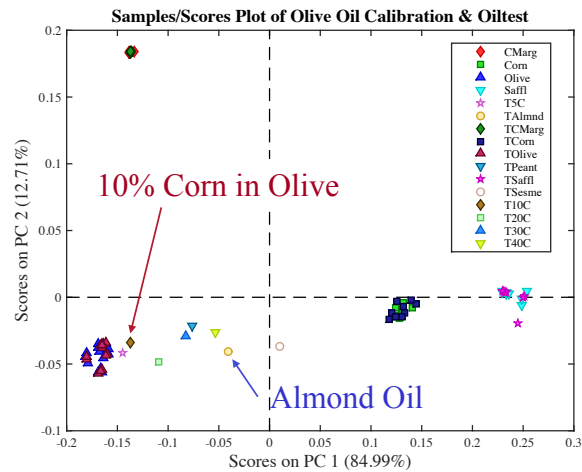
- Bring in the Test Set
 - New samples of Olive, Corn, Corn Marg. & Safflower
 - Other Oils
 - 5, 10, 20, 30 & 40% Corn Oil in Olive Oil
- MSC
- Mean center
- PCA

Let's try it in Solo

112



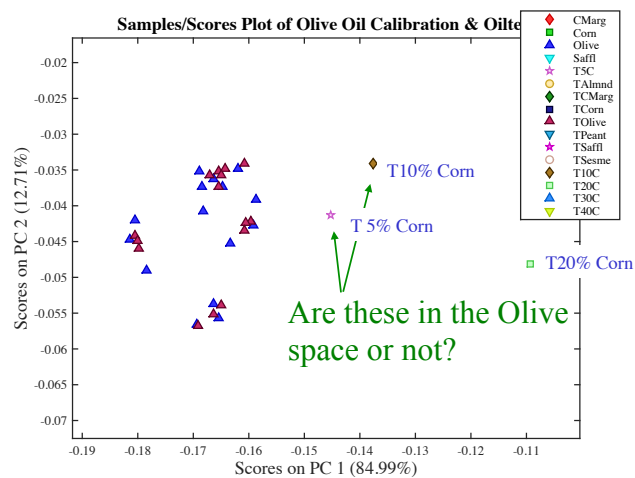
Score Plot Learning Set plus Unknowns “T”



113



Closer View of Olive Oils



114



Return to Slides

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



Outline (Part 1)

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

©Copyright 2004, 2015
Donald B. Dahlberg and Eigenvector Research, Inc.
No part of this material may be photocopied or reproduced in
any form without prior written consent from Eigenvector
Research, Inc. or Donald B. Dahlberg

116



How SIMCA works

- How? →
- 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
 - 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



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

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

Perhaps You Know Your Analytical Technique is Good to $\pm 10\%$

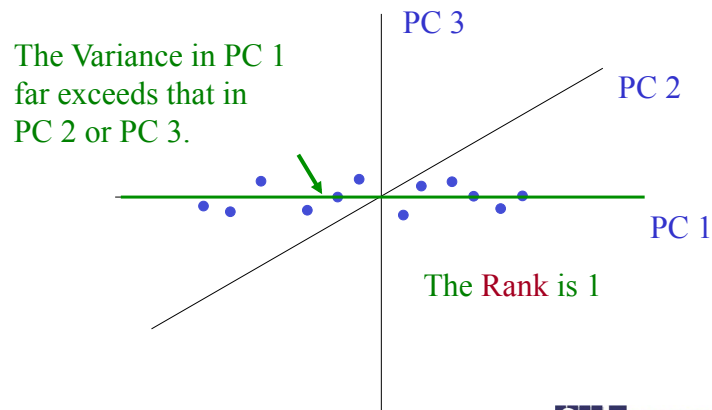
Terminate Model with 3 PCs

Beginning to Model 10% Noise

118



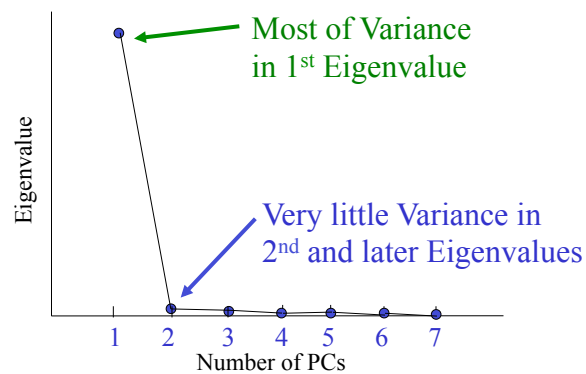
Another Approach is to Examine Distribution of Variance among the PCs



119



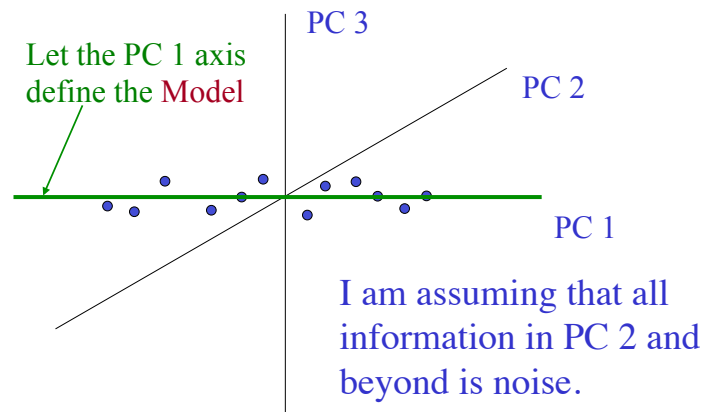
Plot Eigenvalues



120



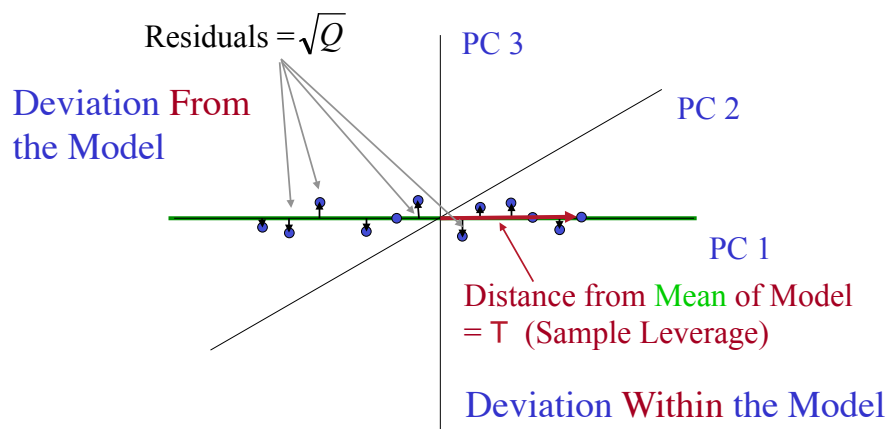
Create One PC Model with Boundaries



121



Individual Points Deviate from the Group in Two Ways



122



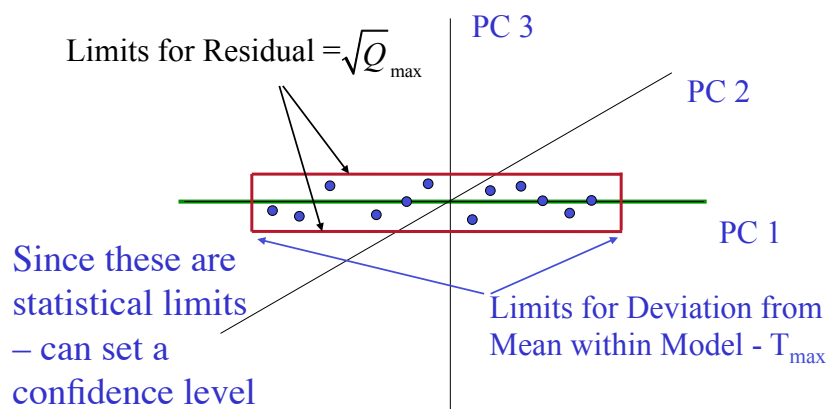
Methods of Measuring Distance

- **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. (this is also how T^2 is calculated)
- **Mahalanobis Distance** has several statistical advantages over **Euclidean Distance**.

123



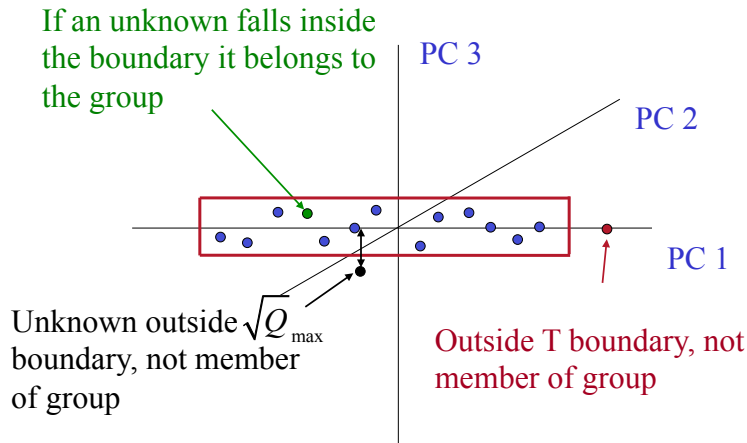
Set Class Boundaries Based on Deviation of Learning Set from Idealized Model



124



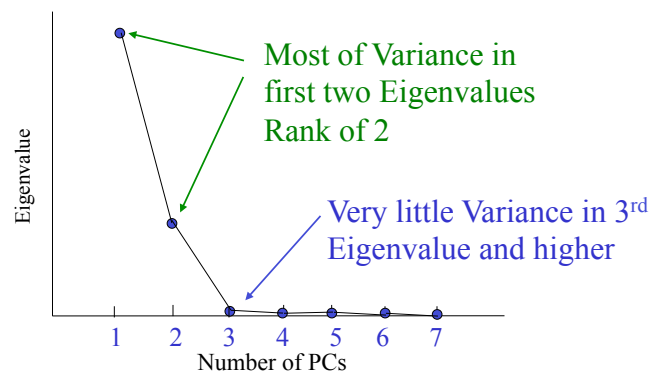
Place Unknown into PC Space



125



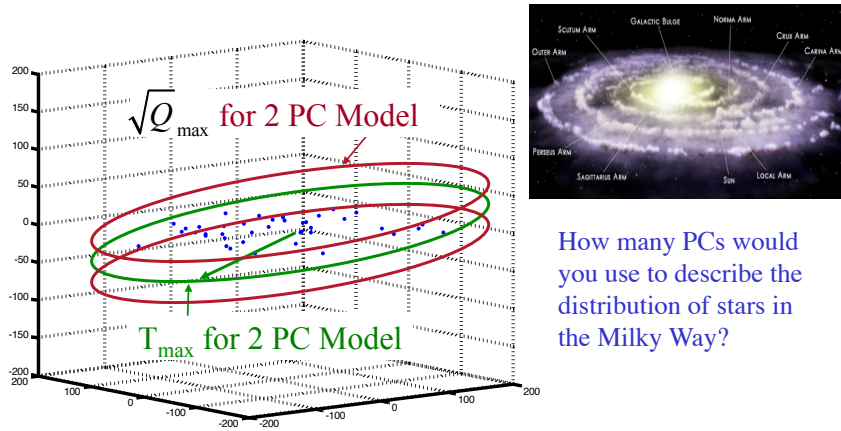
What If the Eigenvalue Plot Looks Like This:



126



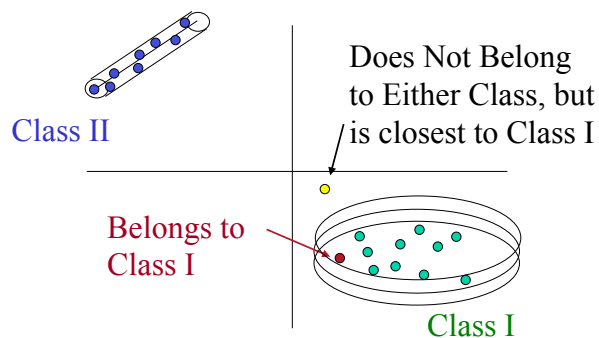
Use Two PC Model



127



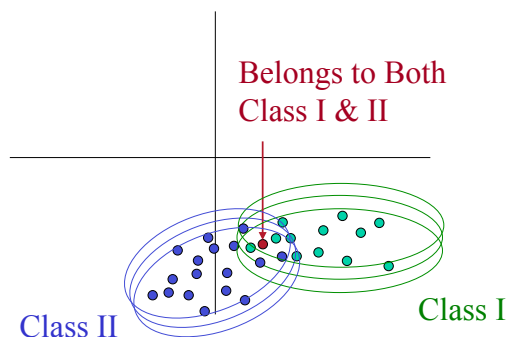
SIMCA tells if a Sample falls within the Boundaries of a Class



128



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



129



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

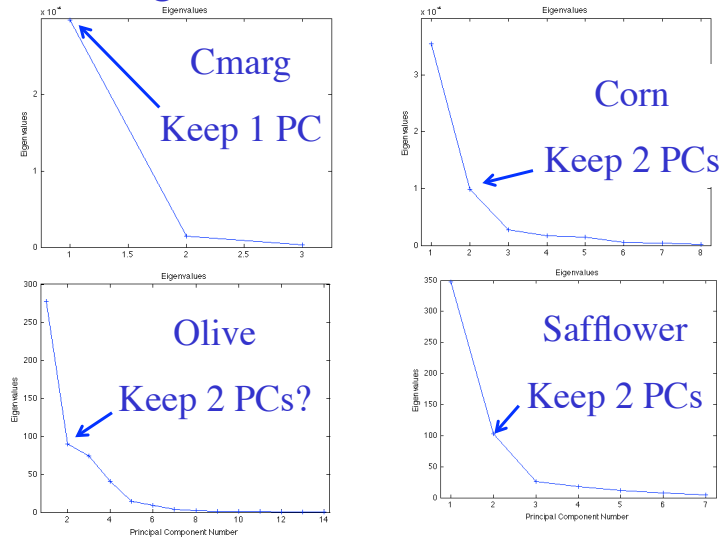
- Outside T only
 - Outside Q only
 - Either Outside T or Outside Q
 - Outside some combined limit *e.g.* $\sqrt{T^2 + Q^2}$
- These picture have been using this rule

Try it in Solo

130



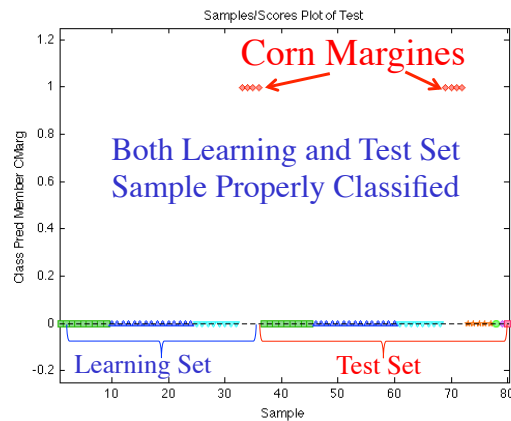
Eigenvalue Plots for Each Class



131



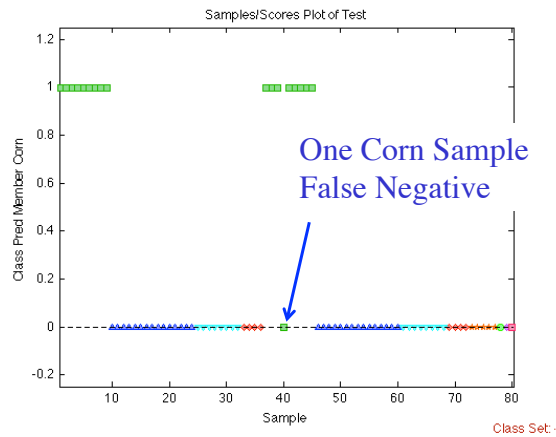
SIMCA Classification of Corn Margarine Samples



132



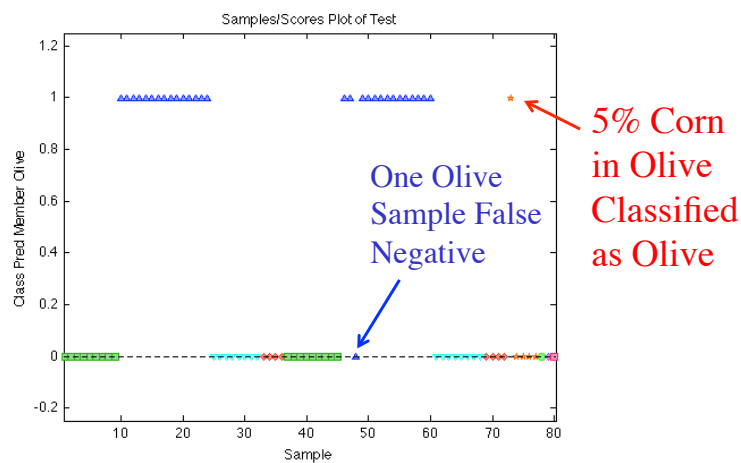
SIMCA Classification of Corn



133



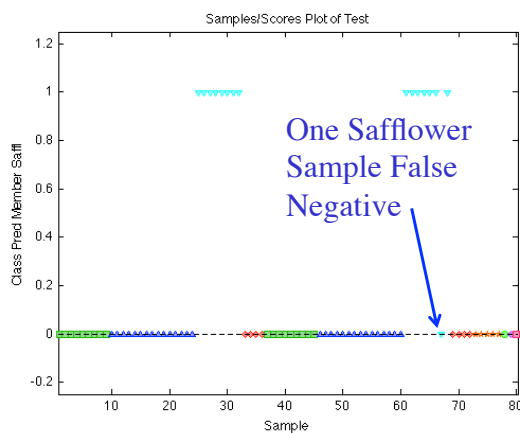
SIMCA Classification of Olive



134



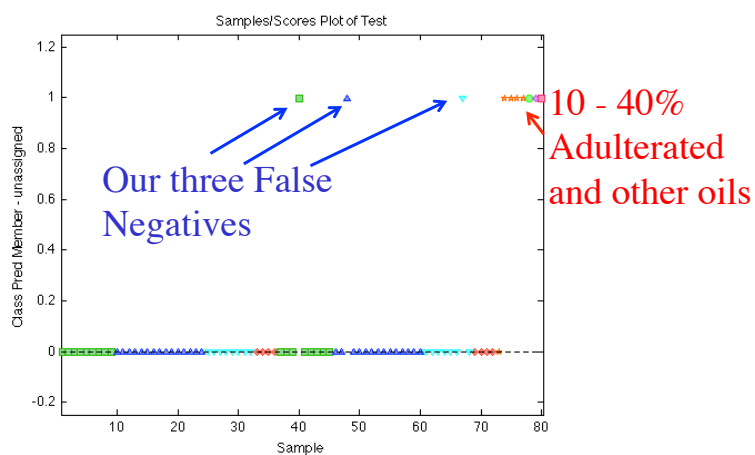
SIMCA Classification of Safflower



135



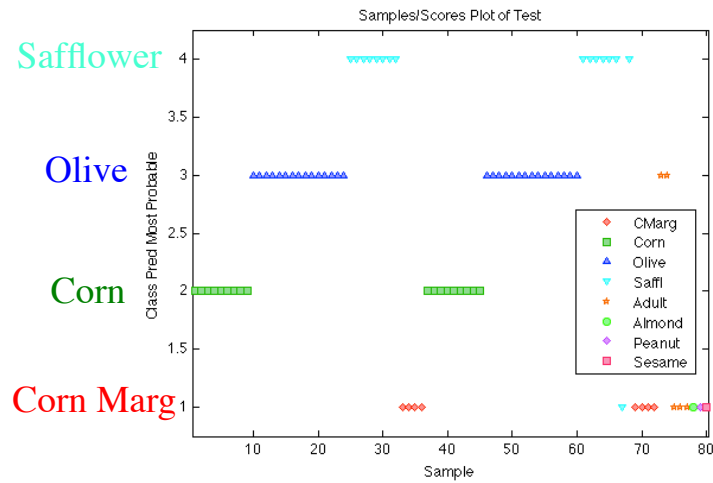
Unassigned Samples



136



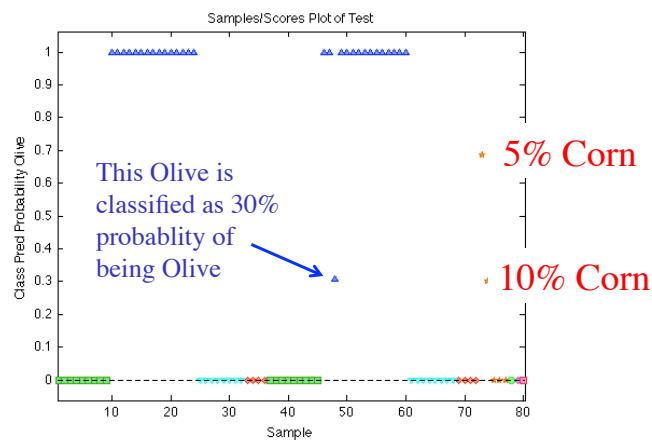
Most Probable – Closest To



137



Probability of Belonging to Olive



138



[Return to Slides](#)

Important Notice: Before Applying PCA or 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.
- Validate Thoroughly With a Well Designed Test Set!
- Models Do Not Last Forever.
- Re-validate Often and Rebuild If Necessary.

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

©Copyright 2004, 2015
Donald B. Dahlberg and Eigenvector Research, Inc.
No part of this material may be photocopied or reproduced in any
form without prior written consent from Eigenvector Research,
Inc. or Donald B. Dahlberg

140



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.

141



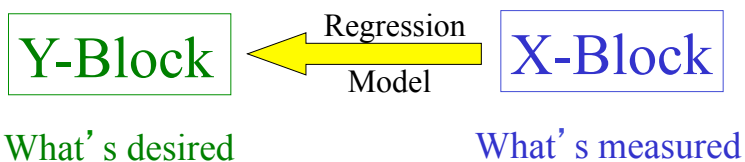
What's Measured:

- 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.

142



Regression



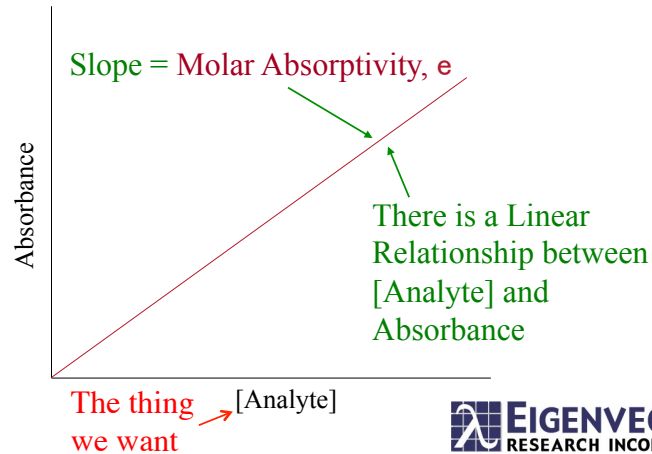
Regression analysis identifies the dependency between two blocks of data.

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

143

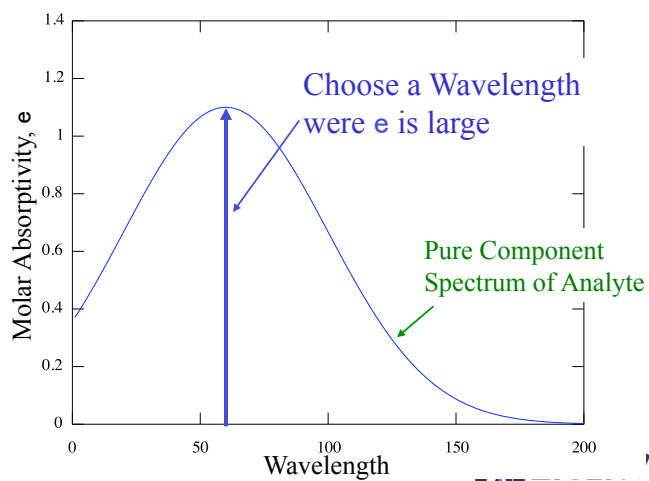


One of the Most Familiar uses of Linear Regression is from the Beer-Lambert Law



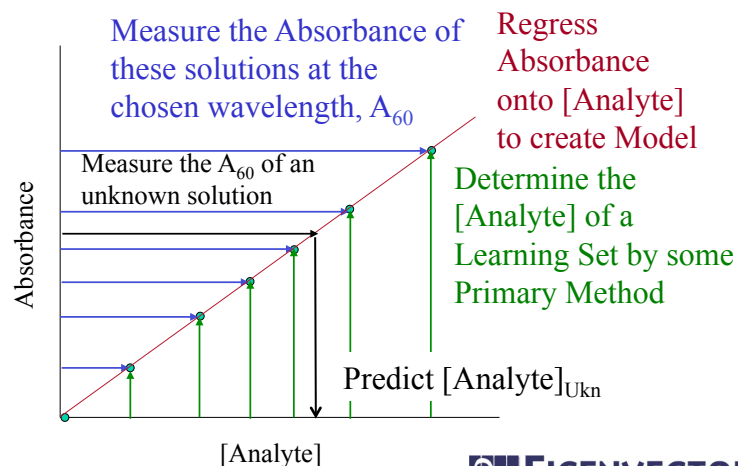
EIGENVECTOR
RESEARCH INCORPORATED

How to Create Beer-Lambert Law Model



EIGENVECTOR
RESEARCH INCORPORATED

Create Beer-Lambert Law Model



146



Outline

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

©Copyright 2004, 2015
Donald B. Dahlberg and Eigenvector Research, Inc.
No part of this material may be photocopied or reproduced in any form without prior written consent from Eigenvector Research, Inc. or Donald B. Dahlberg

147



Sometimes Must Look at an Equation

Classical Least Squares Model = CLS:

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

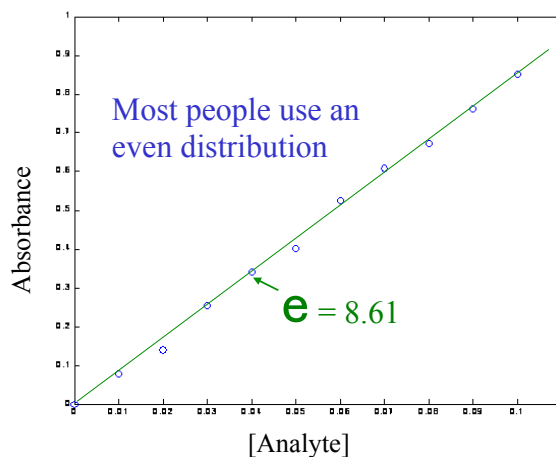
Dependent Variable Independent Variable Regression Coefficient Error in measuring A

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

148



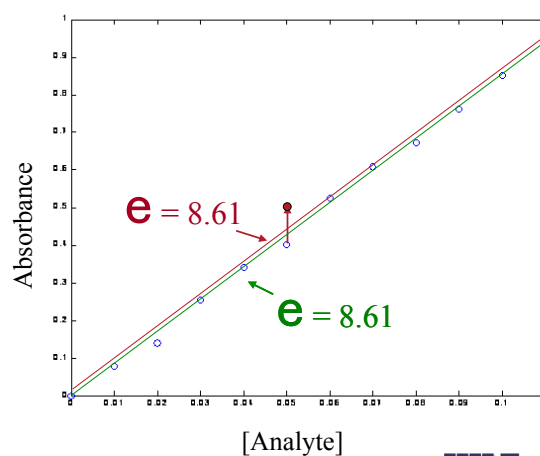
How Should Samples be Distributed?



149



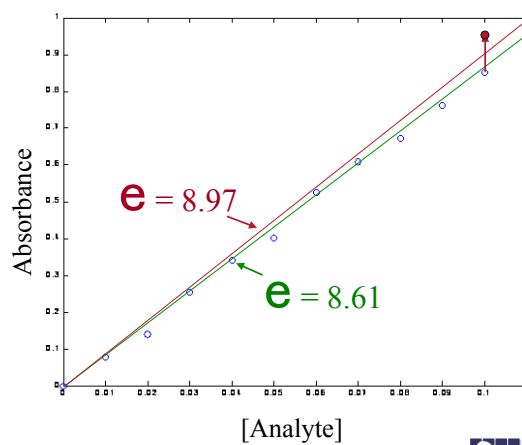
What Happens to the Slope if a Point is Moved Due to Measurement Error?



150



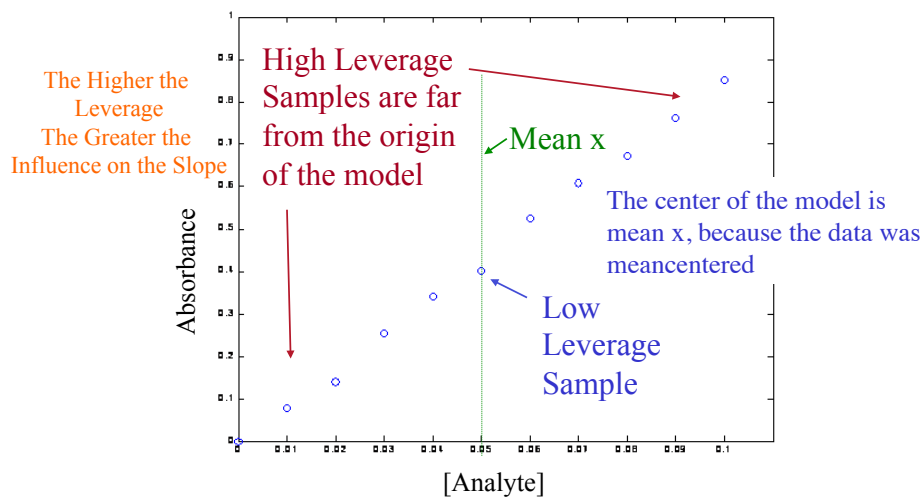
What Happens to the Slope if a Point is Moved Due to Measurement Error?



151



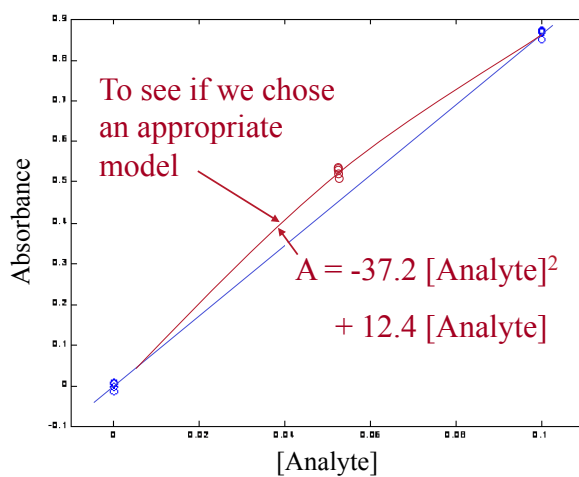
It's About Leverage Again



152



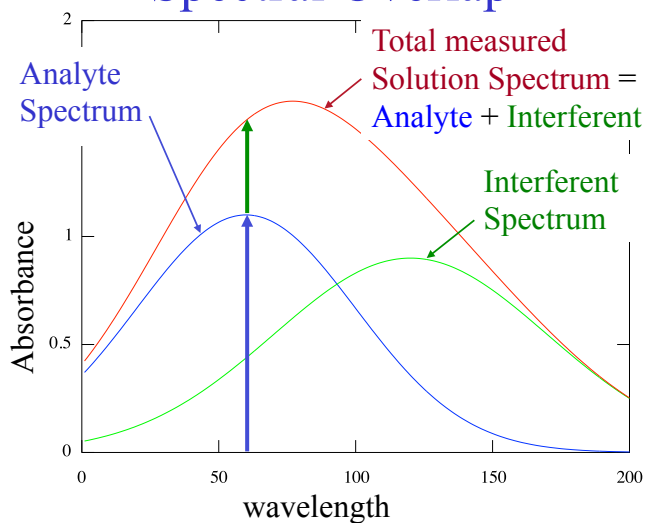
So Why Bother with Middle Points?



153



Spectral Overlap



154



Beer-Lambert Law Now Looks Like:

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

Have two
Concentrations to
worry about

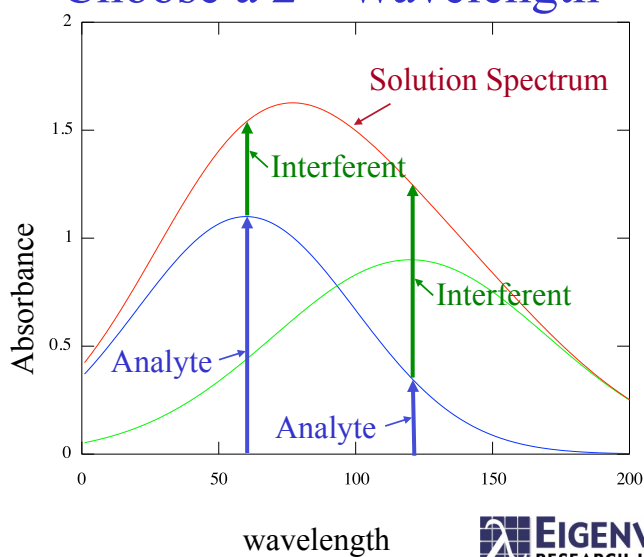
Have two Molar
Absorptivities to
Determine

Two Unknown Constants will
require at least two equations

155



What to Do? Choose a 2nd Wavelength



156

EIGENVECTOR
RESEARCH INCORPORATED

Now Have Two Equations

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

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

And FOUR Unknowns to Determine

157

EIGENVECTOR
RESEARCH INCORPORATED

What to Do?

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

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

Make up a 2nd Solution using different concentrations:

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

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

Solve the equations simultaneously for the four e 's

158



Hey, I Thought This Workshop Was
without Equations!?



159



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.

160



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!

161



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

162



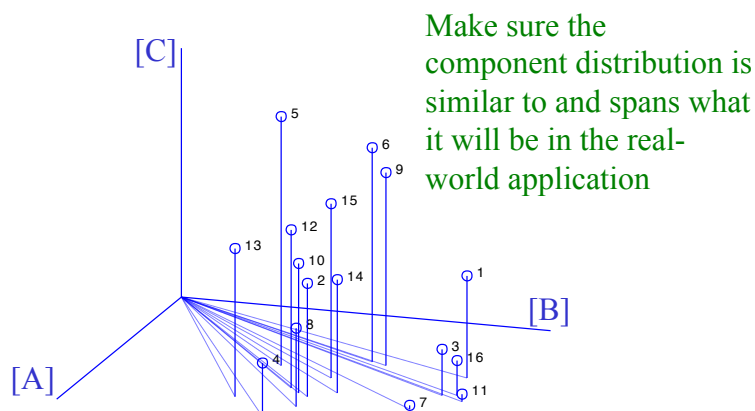
Sixteen Solution Learning Set Three Component System

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

163



Sample Distribution



164



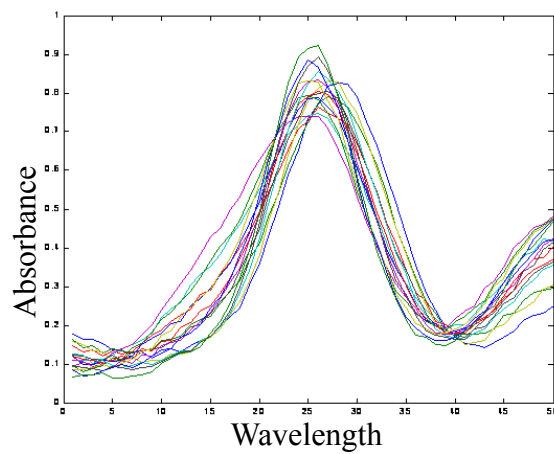
Sample Dilemma

- 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.

165



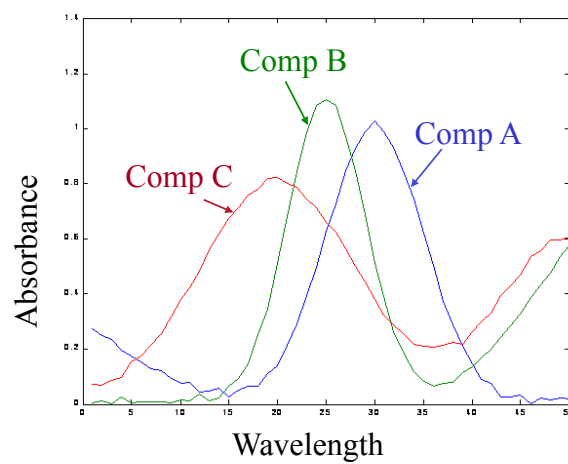
Spectra of 16 Learning Set Spectra



166



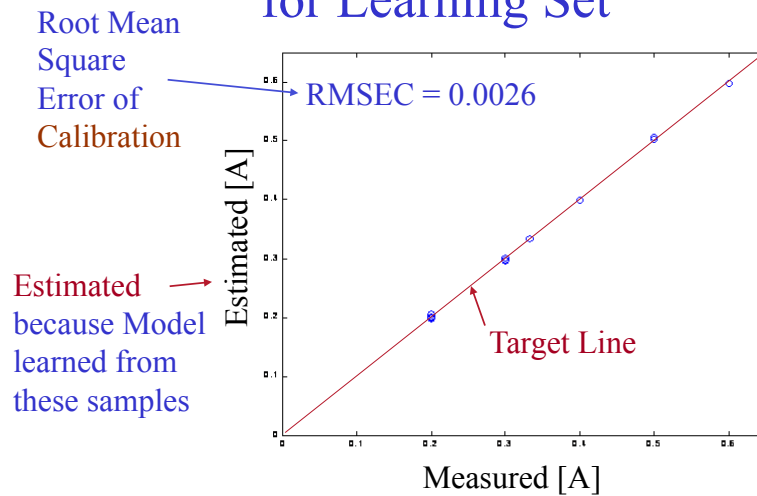
Pure Component Spectra Obtained from CLS



167



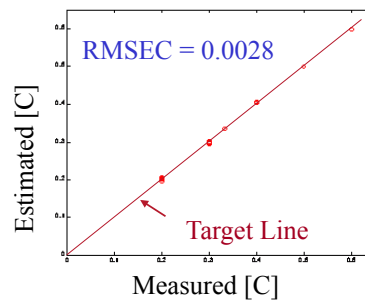
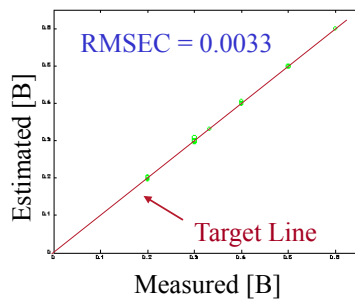
Estimated vs. Measured [A] for Learning Set



168



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



169



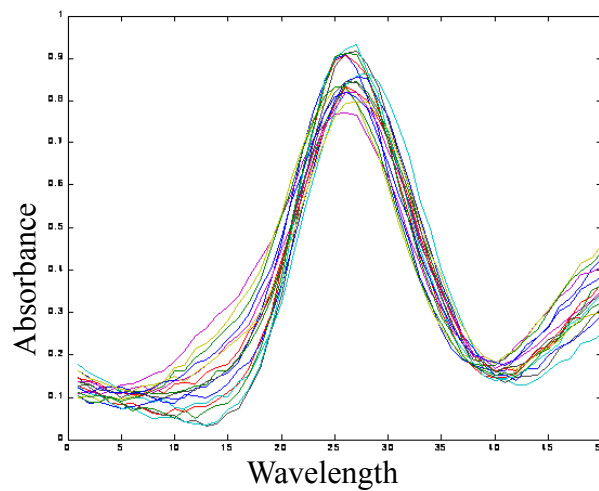
Estimation versus Prediction

- 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**

170



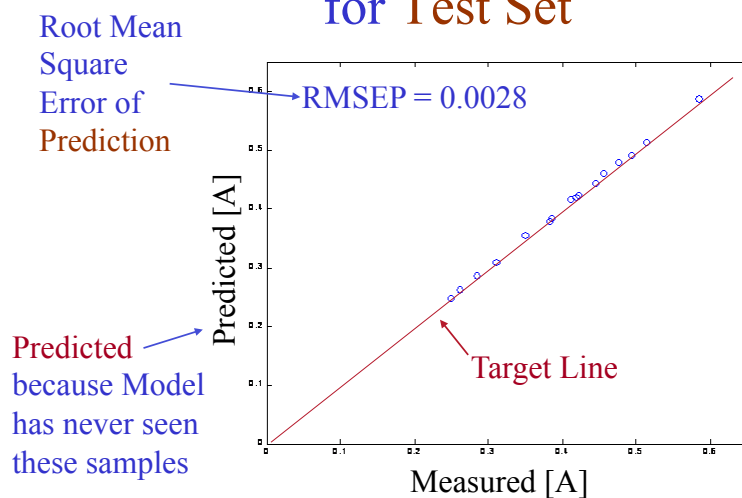
Spectra of 16 Sample Test Set



171



Predicted vs. Measured [A] for Test Set



172



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

173



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

174



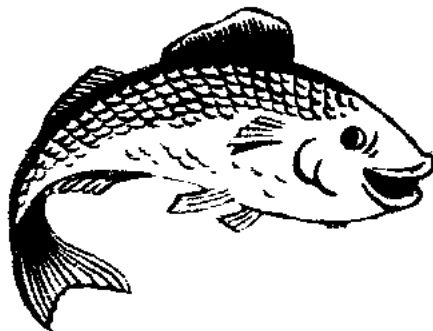
Outline

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

175



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?

176



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

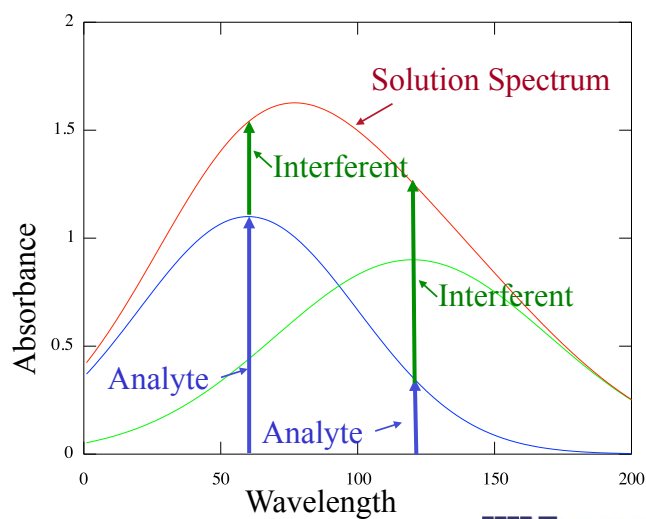
Notice the thing we ultimately want is on the left side of the equal sign.

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

177



Let's Look at MLR Applied to Our

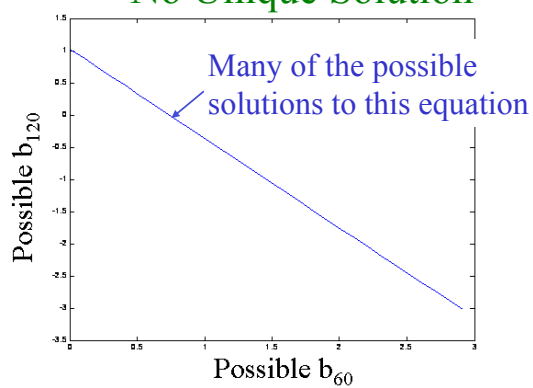


178



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

One Equation - Two Unknowns
No Unique Solution



179



What to Do?

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

Make up a 2nd Solution using different concentrations:

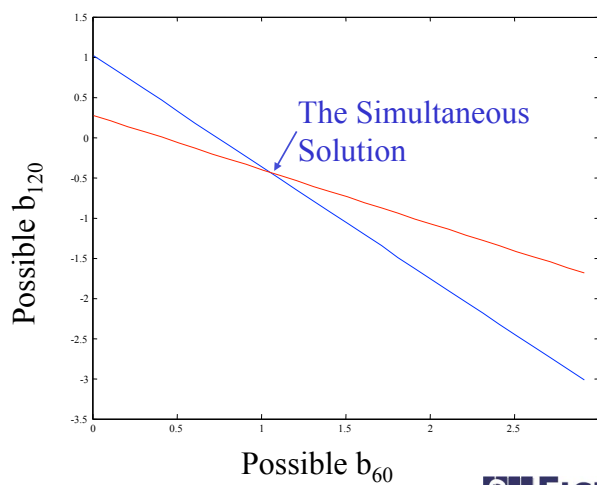
$$[\text{Analyte}]_{\text{II}} = A_{60,\text{II}} b_{60} + A_{120,\text{II}} b_{120}$$

Notice the other interfering chromophores are not in the equation.

180



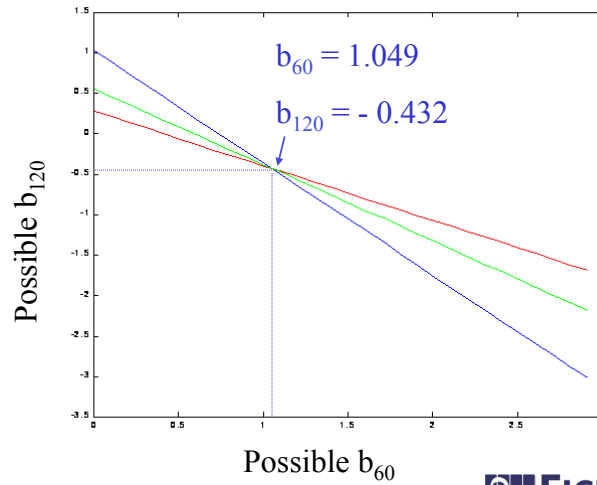
Two Equations - Two Unknowns



181



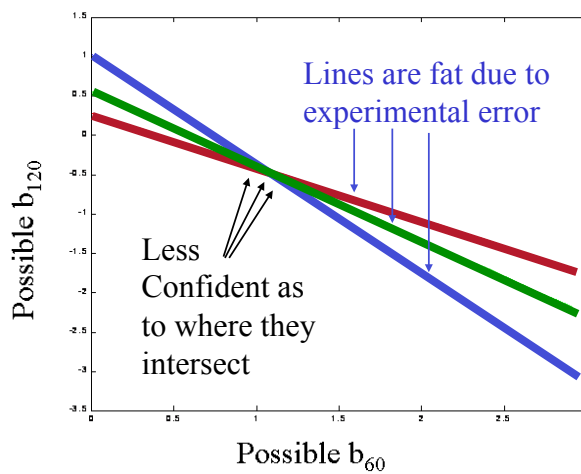
Add a Third Experiment Overdetermined Solution



182



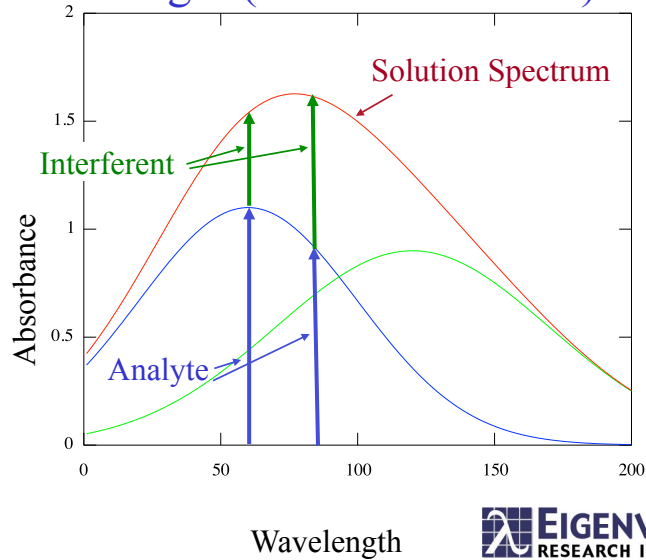
Problem Solved? Not Exactly!



183



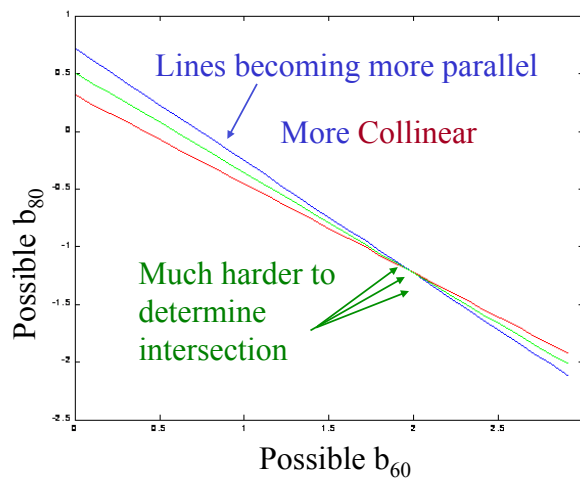
What if We Choose a Different 2nd
Wavelength (closer to the first)?



184



Using Wavelength 60 & 80



185



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

186



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?

187



Outline

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

©Copyright 2004, 2015
Donald B. Dahlberg and Eigenvector Research, Inc.
No part of this material may be photocopied or reproduced in any
form without prior written consent from Eigenvector Research,
Inc. or Donald B. Dahlberg

188



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

189



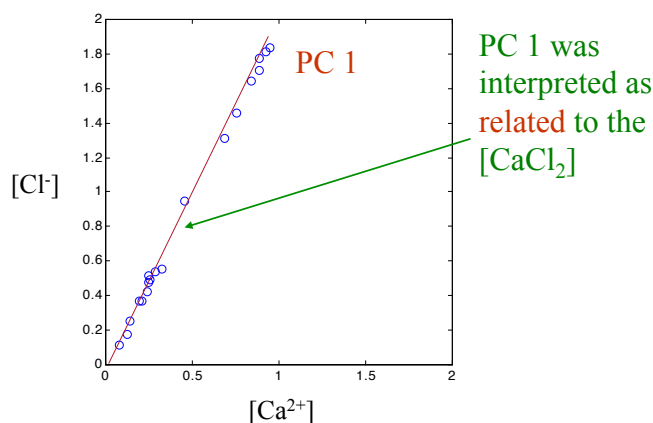
PCA

- 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 \leftrightarrow Variables
- Scores are the positions of the samples in the new PC space.
 - Scores \leftrightarrow Samples
- Perhaps MLR on the PCA scores?

190



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

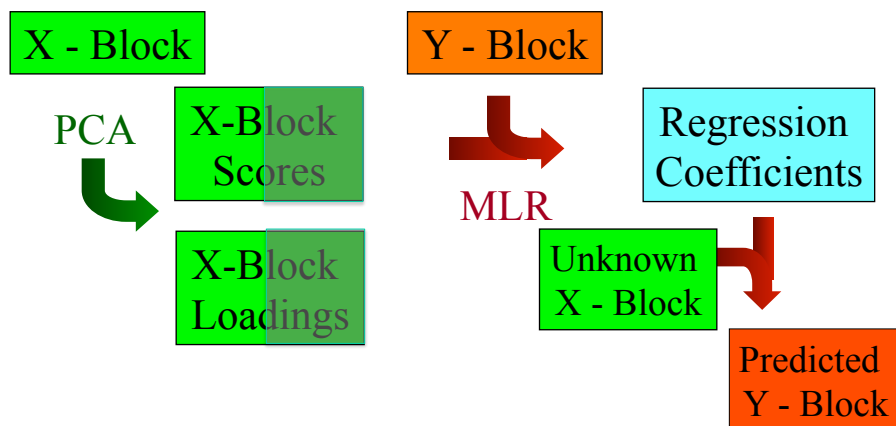


191



Principal Components Regression

PCR

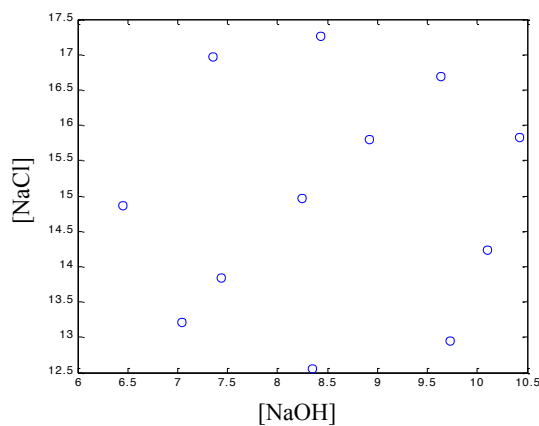


192



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

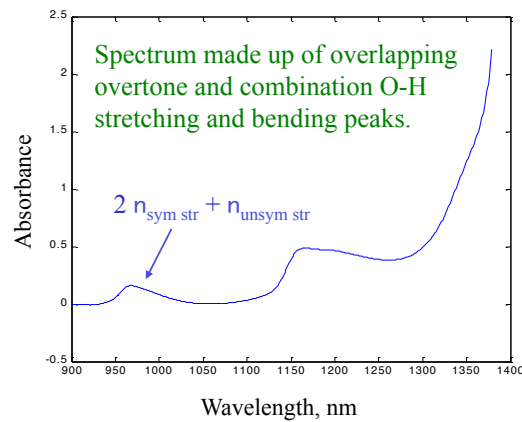
Have 12 Solutions of NaOH and NaCl in Water



193



Typical SW-NIR Spectrum of Caustic Brine Solution



194



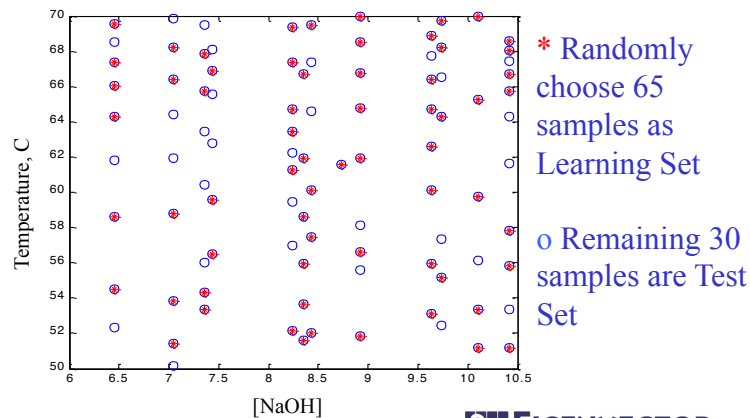
Peak Shift

- The NIR 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

195



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



196

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



PCR on Learning Set

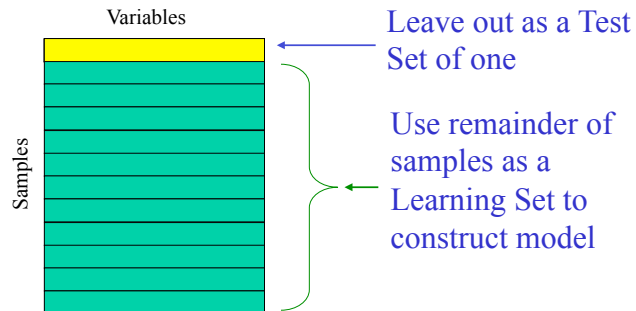
- Should we mean center or autoscale?
 - Usually just mean center spectra
 - recall that this means that we're allowing for an intercept other than zero
- How many PCs to keep in the model?
 - Cross-validation

197



Many Patterns of Cross-Validation

- Leave-One-Out Cross-Validation



198



Leave One Out Cross-Validation

Calculate Errors of Prediction for these Models

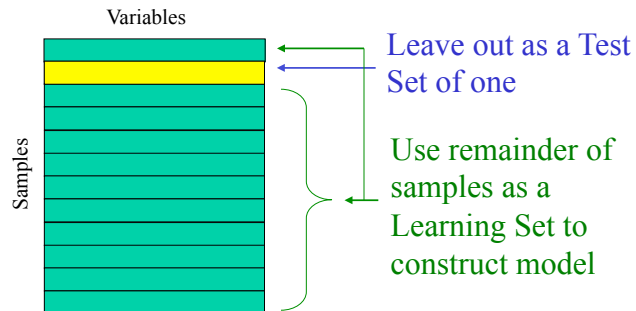
Sample left out	Error of Prediction using					
	1	2	3	4	5	... N PCs
1	—	—	—	—	—	...

199



Many Patterns of Cross-Validation

- Leave-One-Out Cross-Validation

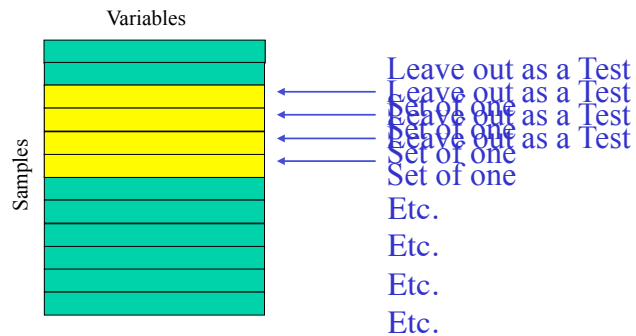


200



Many Patterns of Cross-Validation

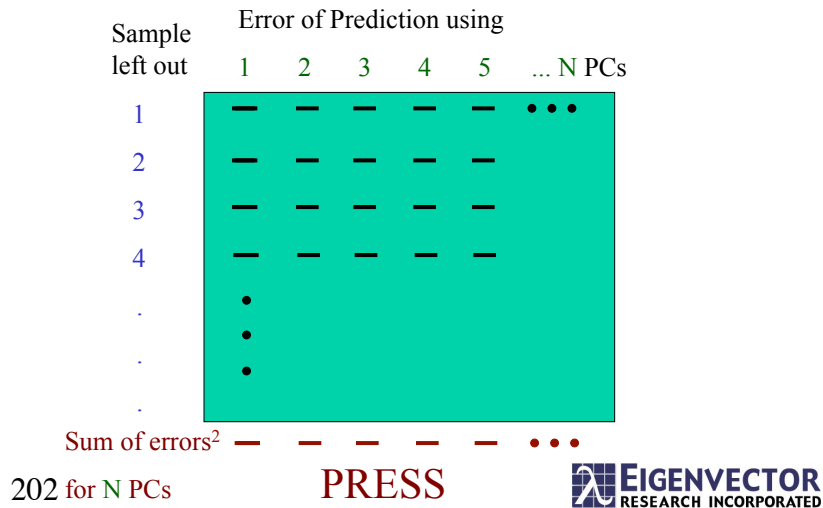
- Leave-One-Out Cross-Validation



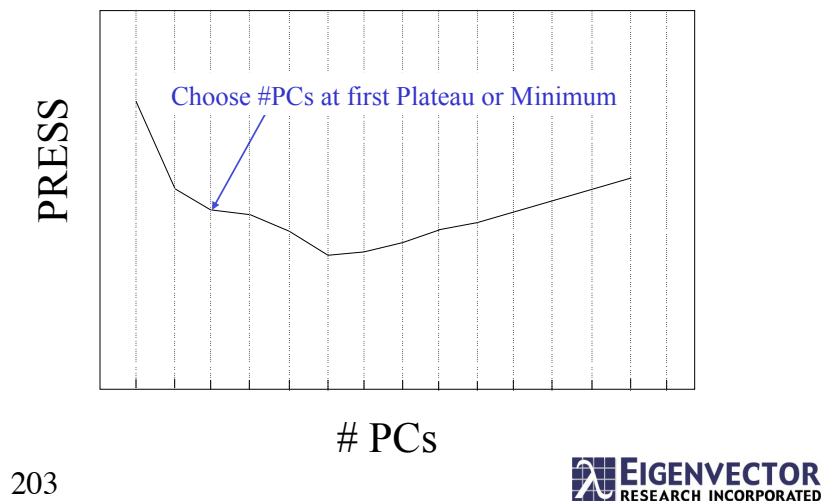
201



Leave One Out Cross-Validation

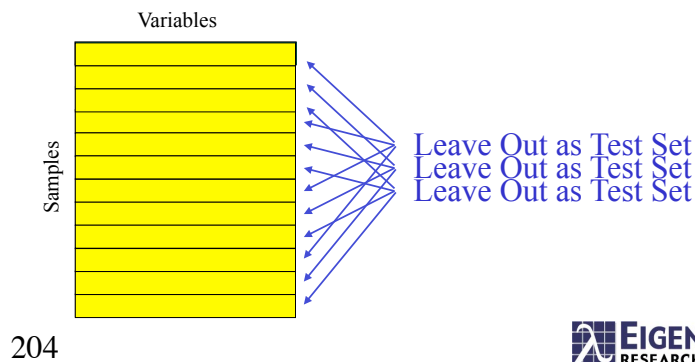


Construct PRESS Plot



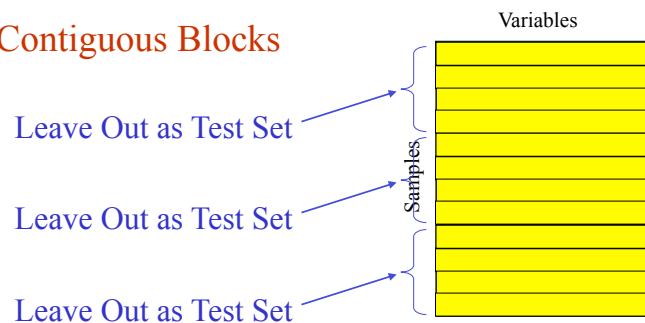
Many Patterns of Cross-Validation

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



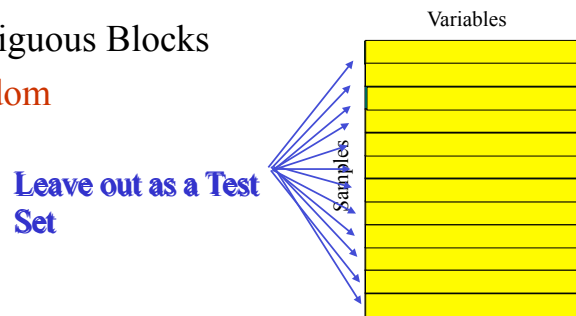
Many Patterns of Cross-Validation

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



Many Patterns of Cross-Validation

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



206



Cross-Validation Rules of Thumb

- 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)

207



Cross-Validation for PCR Example

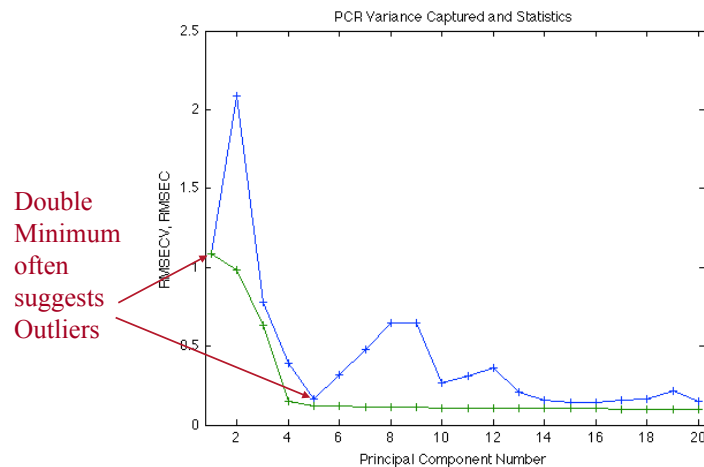
- Venetian Blind
- 8 Splits (approx. ~65 samples)
- Calculate RMSECV for up to 20 PCs
 - expect the number to be <20

Let's try this in Solo

208



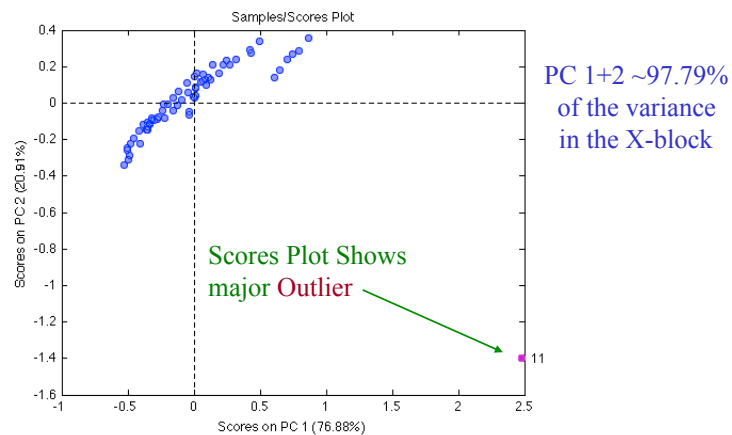
PRESS Plot



209



Keep 5 PCs
just to see what is going on

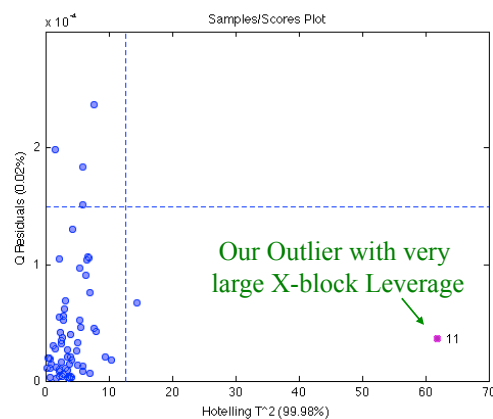


210



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

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

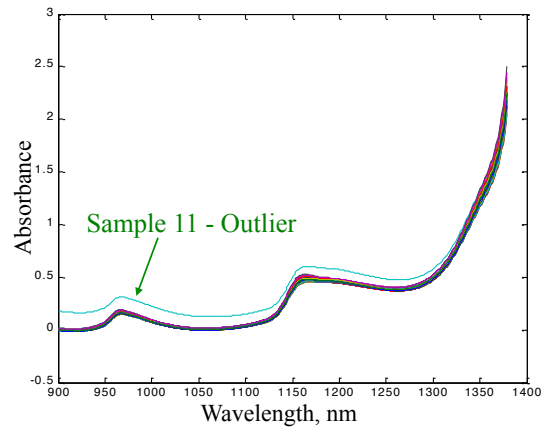


It appears to be within the model, because it pulls the model to itself

211



Spectra of Learning Set



212



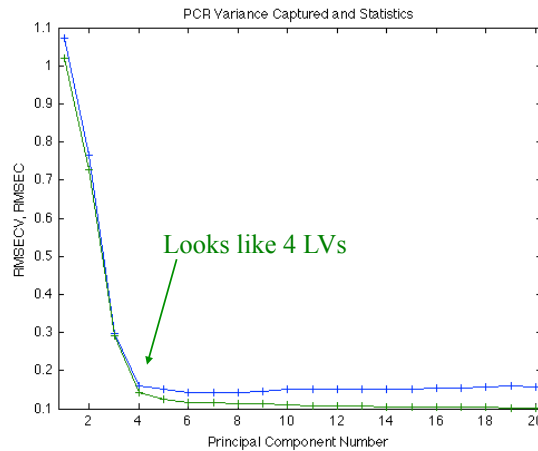
Important Lessons

- 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 **T²** for Outliers before building calibration model
 - Exploratory data analysis

213



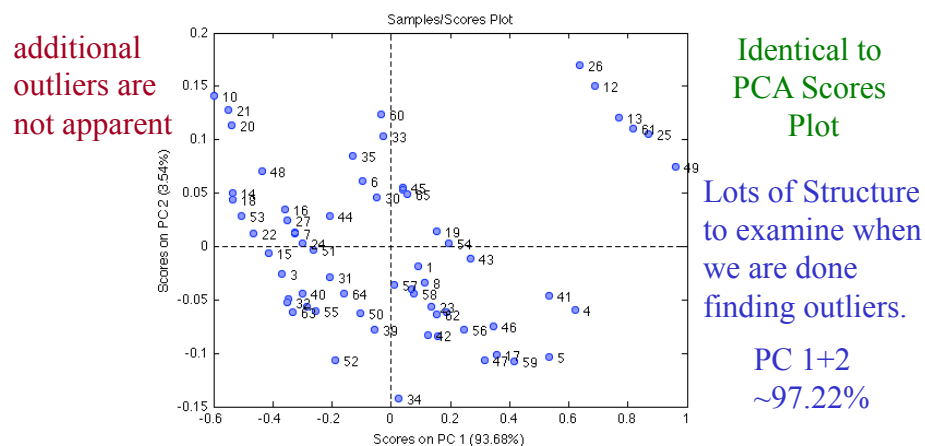
Delete Outlier Sample 11, Rebuild Model and Examine PRESS Plot



214



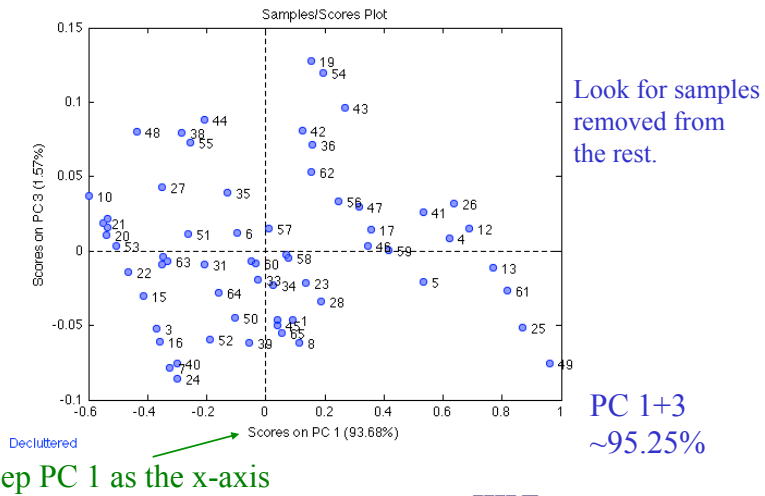
Scores Plot



215



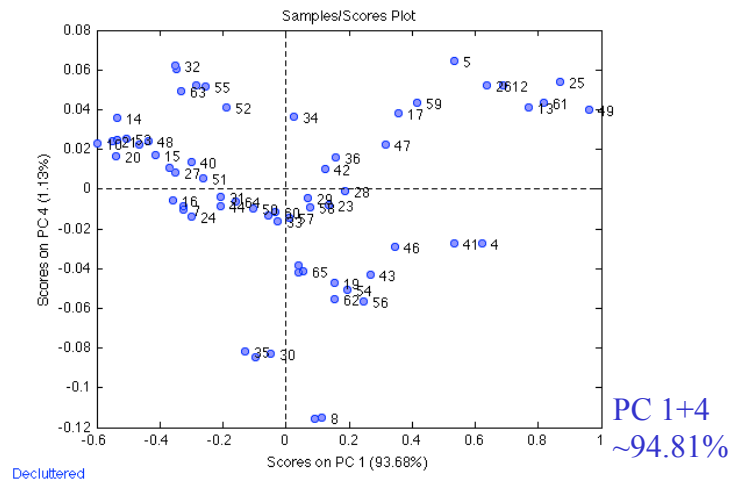
Examine All the PC Scores



216



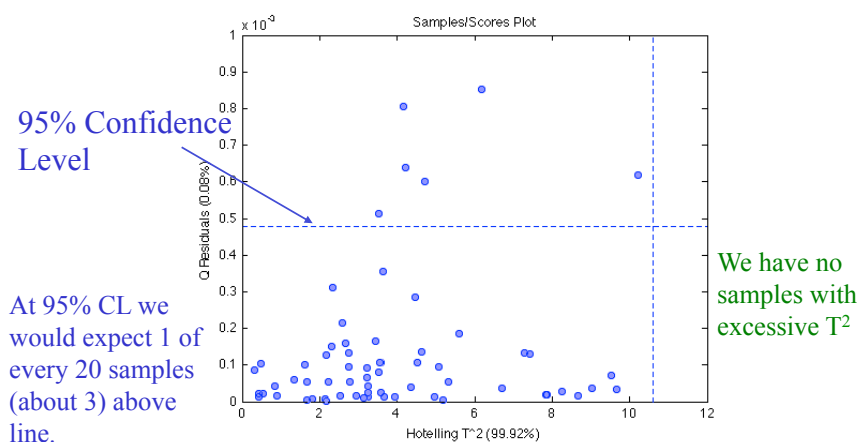
Scores PC 4 vs PC 1



217



X-Block Influence Plot



218



Looking for X-block Outliers

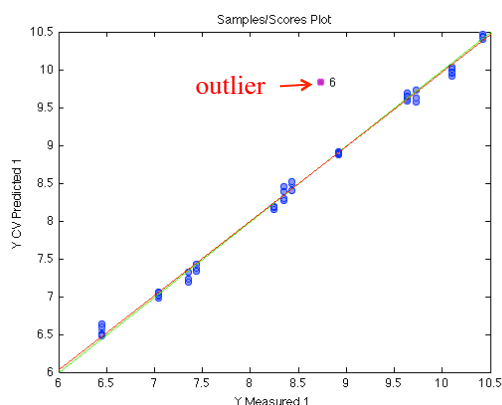
- Look for **isolated** sample(s) in Scores Plots.
- Look for samples with **excessive** Q Residual or Hotelling's T².
- An **Outlier** is not necessarily a bad sample, just unique. Examine it:
 - If good, add more like it to Learning Set.
 - If bad, fix it or remove it.

219



How Did We Do?

Estimated [NaOH] vs. Measured [NaOH]



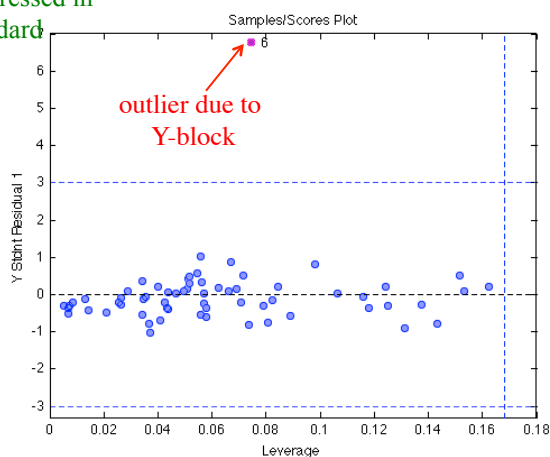
Deviations from the diagonal are used to calculate RMSECV. This plot is for the Learning Set. If plotting fit it should read “estimated” instead of “predicted.” Here, we’re plotting CV predictions.

220



Model or Y-Block Influence Plot

Residual expressed in units of Standard Deviation



Large Leverage → implies error in X-block

Model Leverage

221



Return to Slides

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.

222



Remove or Fix Sample #6

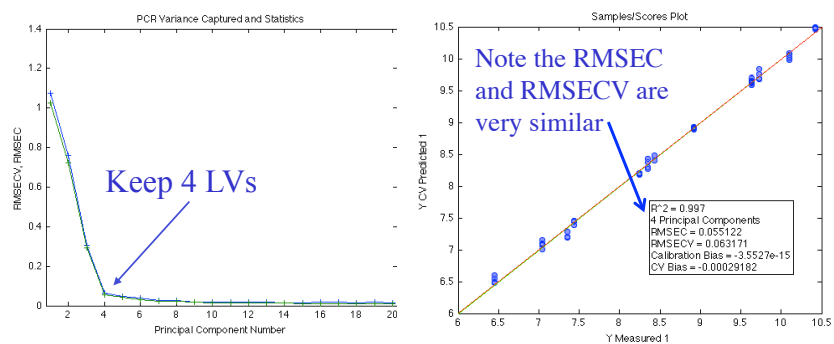
- It turns out that there was an “entry error” for the “known” [NaOH] for Sample #6
 - The measured was 8.7325
 - Entered in the data was 9.7325
- Once corrected, we can make another model
- Check for outlier again using Influence & Scores Plots
- Modeling tends to be iterative
 - Use what is learned at each step to help in subsequent models

223

Go back to Solo



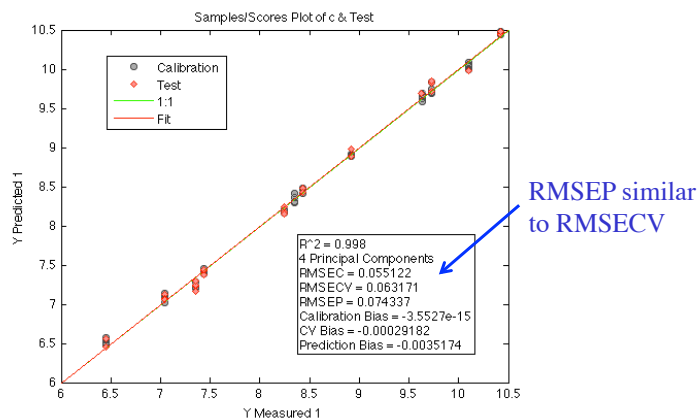
PRESS and Prediction Plot Plots



224



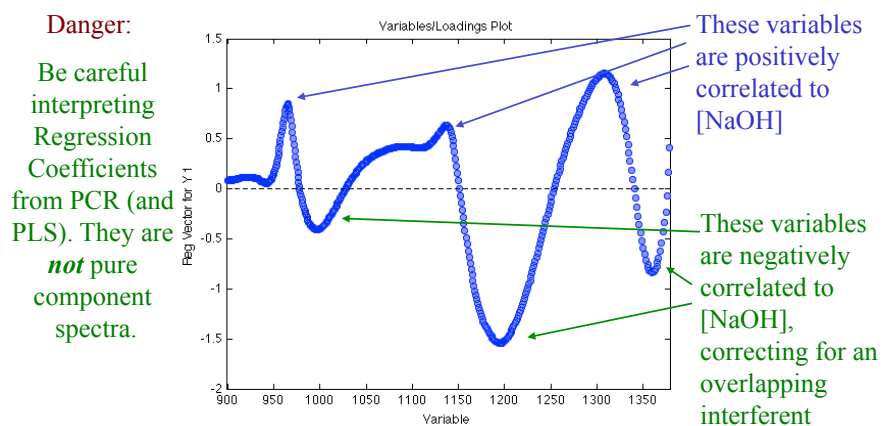
Apply Model to Test Set



225



This Is the Model Regression Coefficients

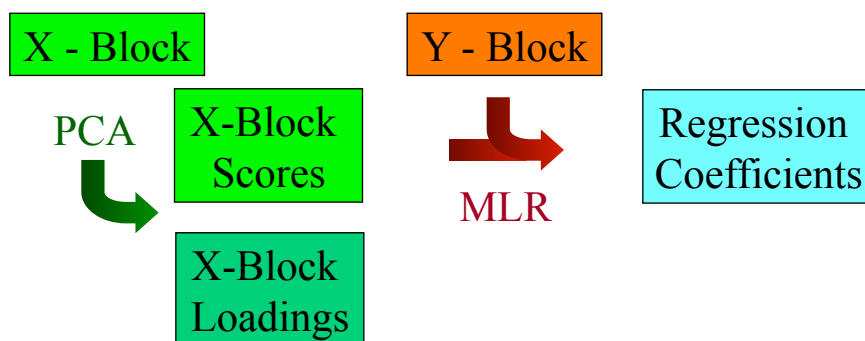


226



Return to Slides

Recall How PCR Works



227



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 \leftrightarrow Latent Variables?

228



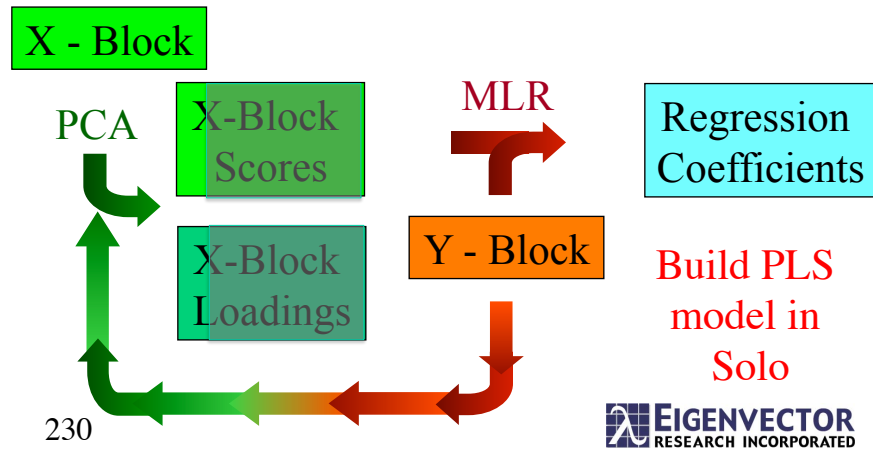
Outline

- 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

229

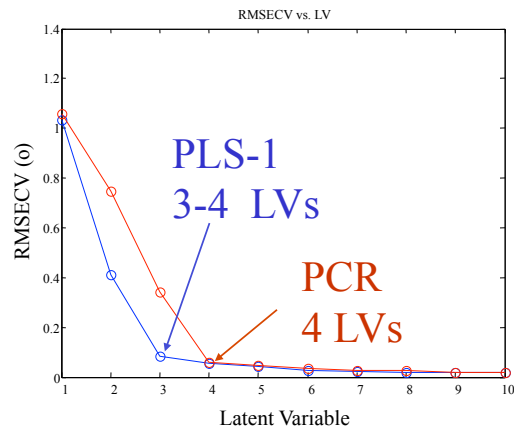


Modification to PCR Partial Least Squares, PLS-1



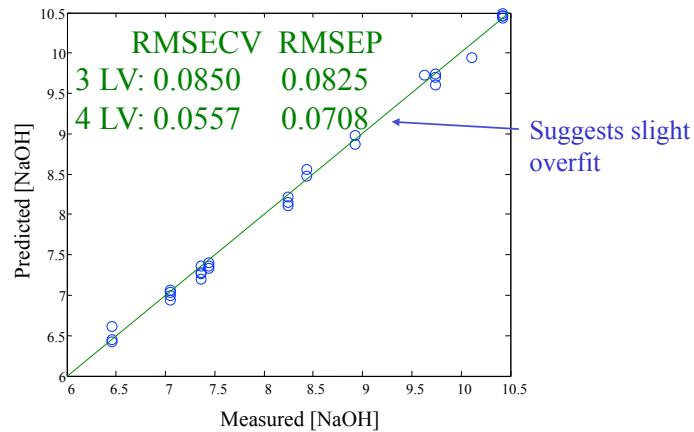
Return to Slides

Compare PRESS Plots



231

PLS-1 Prediction of Test Set with 3 LVs



232



PLS-1 Compared to PCR

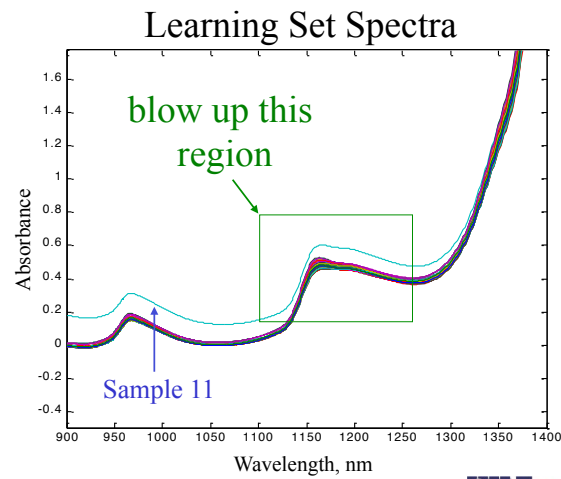
- The hope is that PLS-1 will require fewer latent variables than PCR requires principal components which suggests that there will be less noise
- The two methods often provide very similar results
- PLS usually brings more useful information into the earlier LVs for easier interpretation

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

233



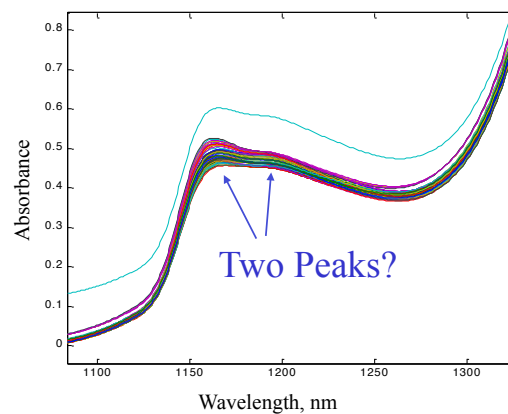
Can We Do Better?



234



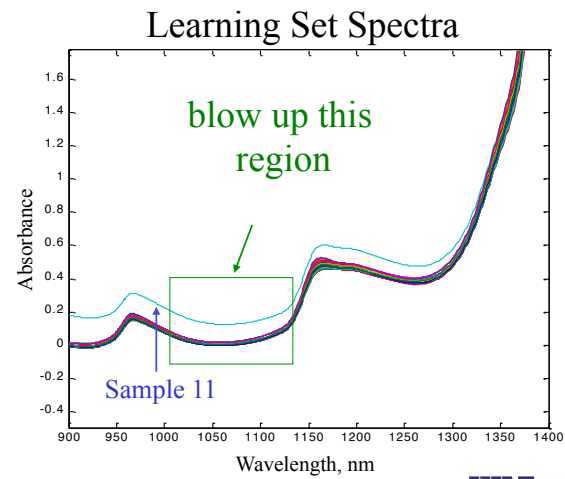
NIR Spectra Composed of Many Broad Overlapping Peaks



235



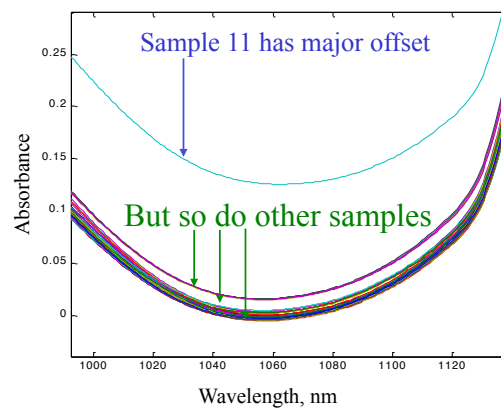
Baseline Offset?



236



Yes

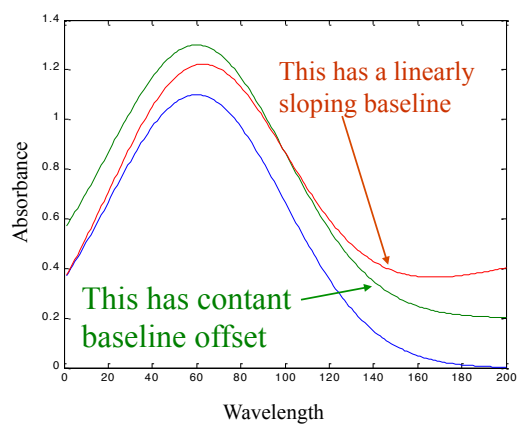


237



Baseline Problems Example

Here are three identical spectra, except:

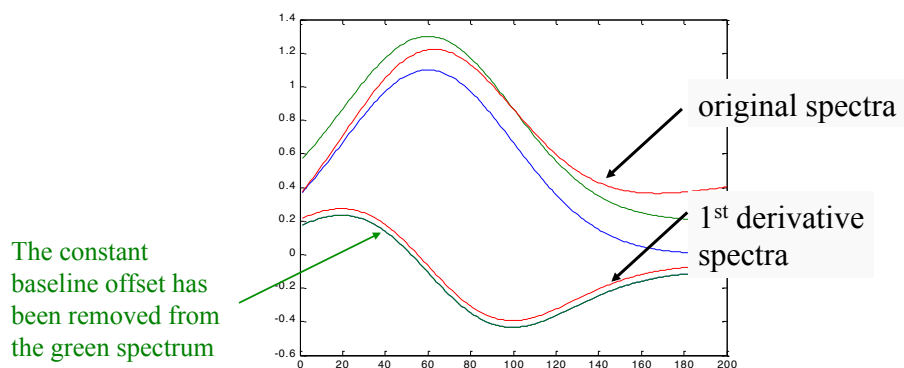


238



Using Derivative Spectra

Take 1st derivative of the Three Spectra

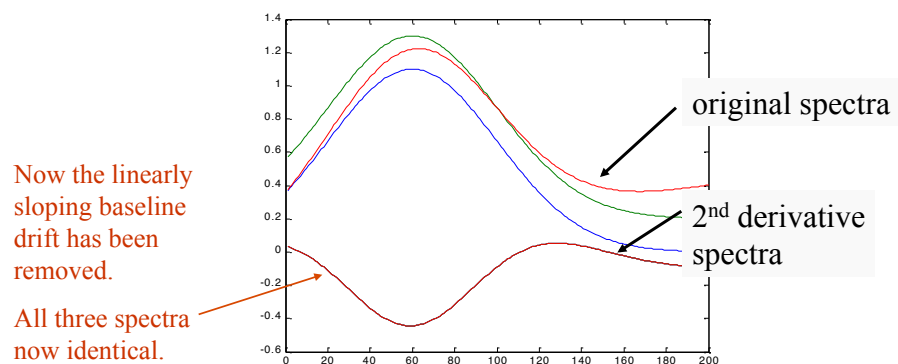


239



Using Derivative Spectra

Take 2nd derivative of the Three Spectra



240



Comments on 2nd Derivative

2nd derivative spectra look different

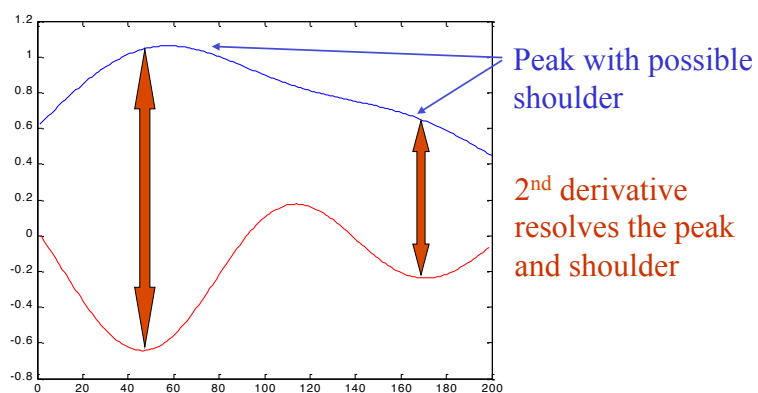
The peak
“maximum” is
now negative.

241



Comments on 2nd Derivative

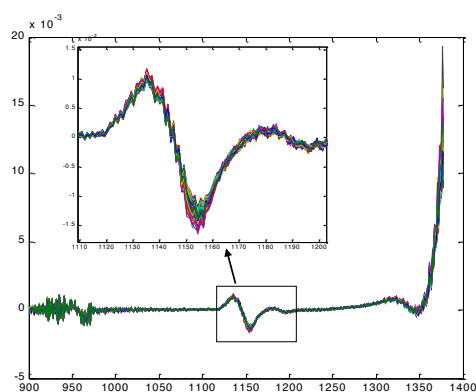
2nd derivative increase resolution



242



2nd Derivative of Learning Set Spectra

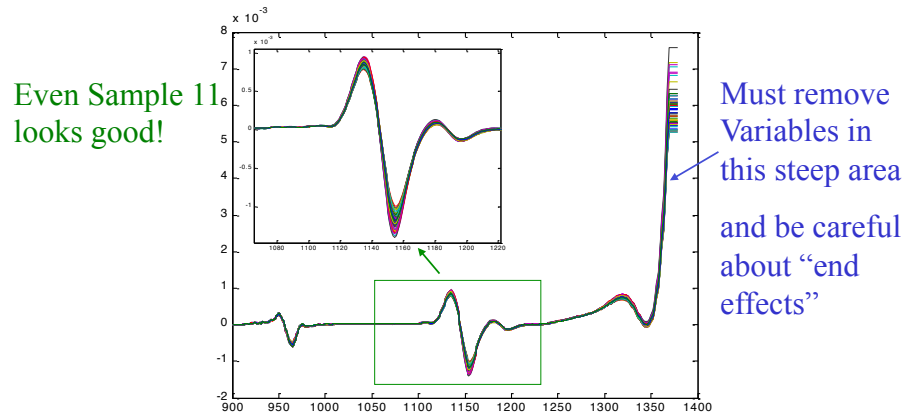


Derivatives can also increase the effect of Noise!

243



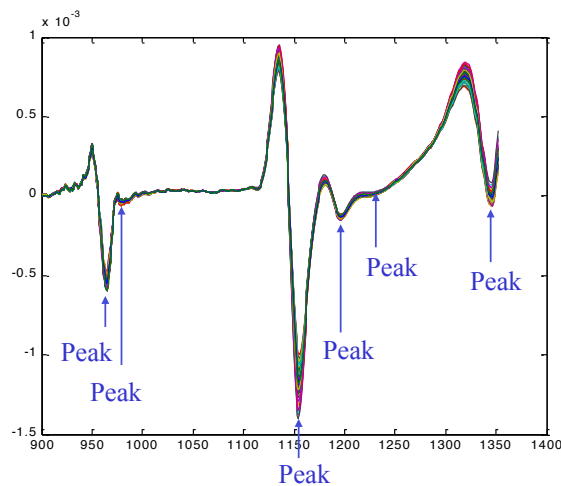
Filter the Data Before Taking 2nd Derivative: Savitzky-Golay



244



Notice the Resolution

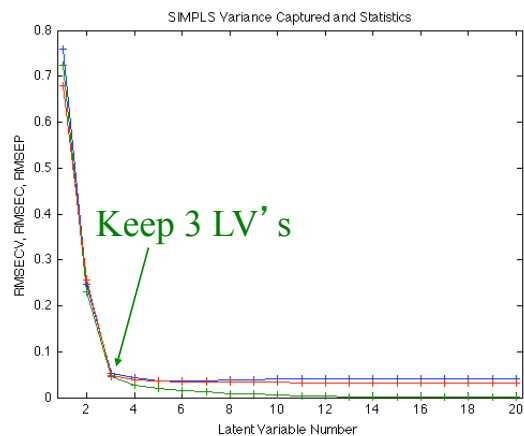


245

Go to Solo



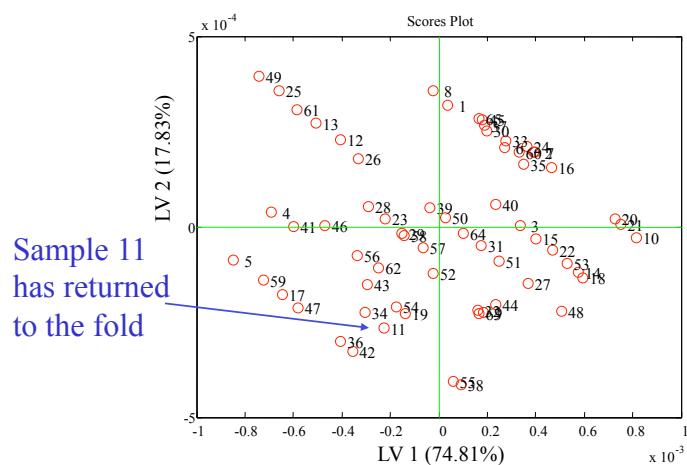
PLS-1 of 2nd Derivative Spectra for the Learning Set



246



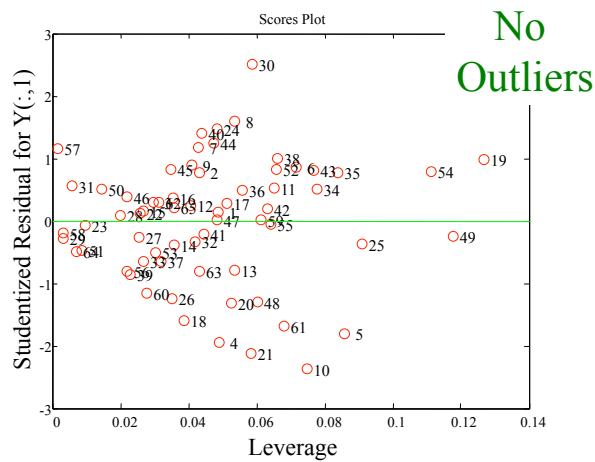
Scores Plot



247



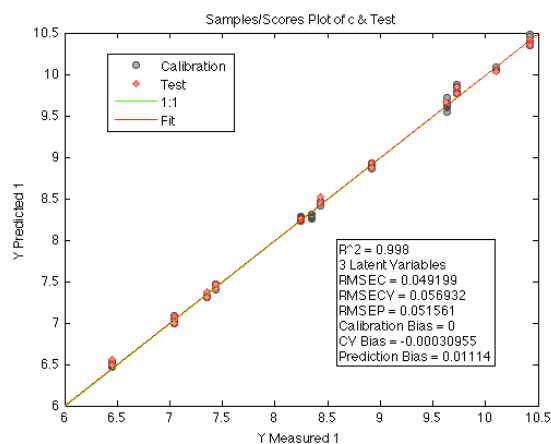
Model Influence Plot



248



Prediction of [NaOH]



249



[Return to Slides](#)

Comparison of Regression Methods and Pretreatments

	RMSECV	RMSEP
PLS-1 2 nd 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

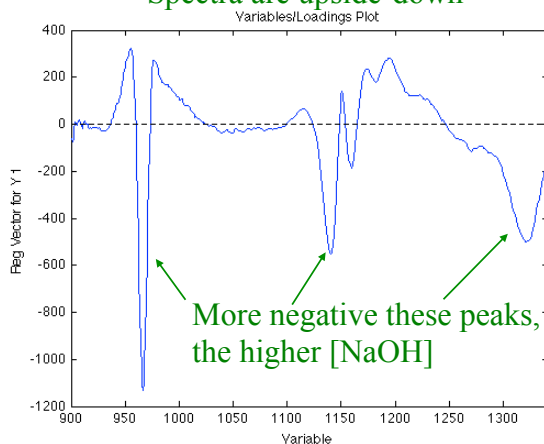
Need to ask: is the difference statistically significant
with respect to error in the reference method.

250



Regression Coefficients

Remember 2nd Derivative
Spectra are upside-down



251



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

252



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

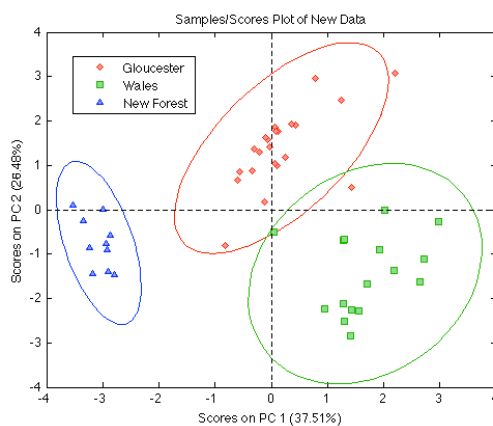
253

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



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

Not very good discrimination.



254



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

↖ Gloucester (yes or no?)
 ↖ Wales (yes or no?)

255

Try it in Solo



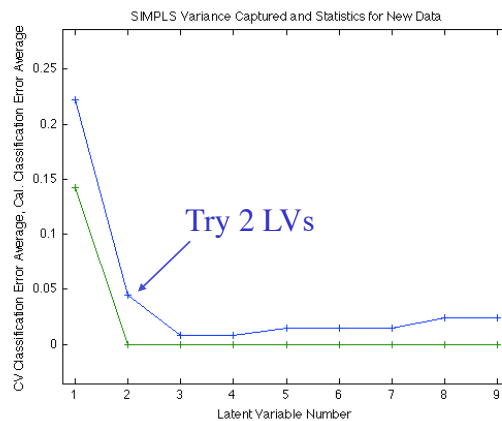
Split Samples into a Learning Set (32 samples) and Test Set (16 samples) using Kennard-Stone

Learning Set:	Gloucester	12	2	3	Need to balance sets
	Wales	13			
	New Forest	10			
Test Set:	Gloucester	10	2	3	
	Wales	3			
	New Forest	3			

256



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

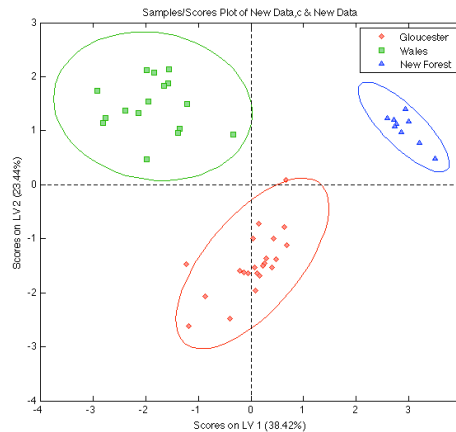


257



PLS-DA Scores PV2 vs LV1

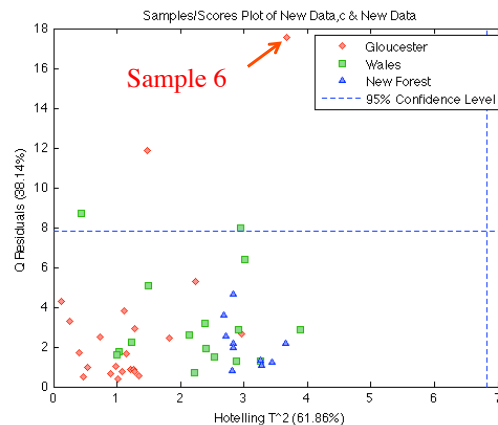
Bringing in the
Y-block during
the creation of
the LVs
improved class
discrimination



258



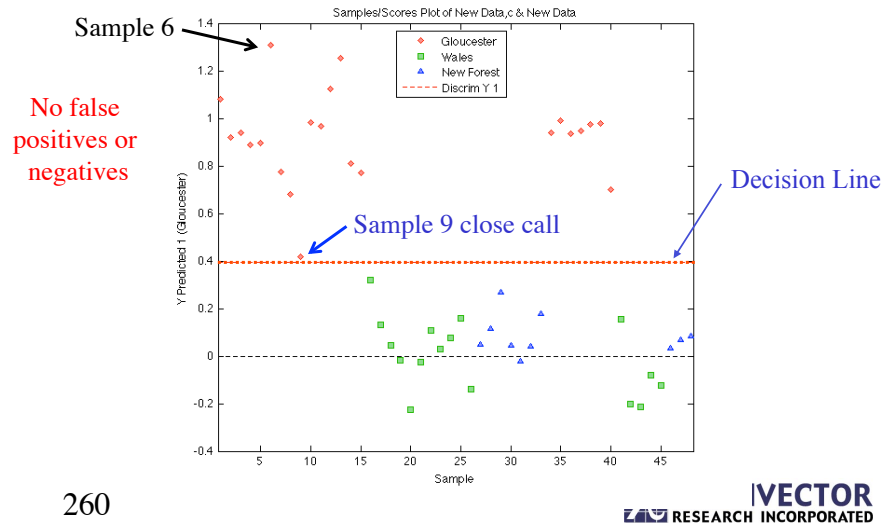
Always Check Influence Plot



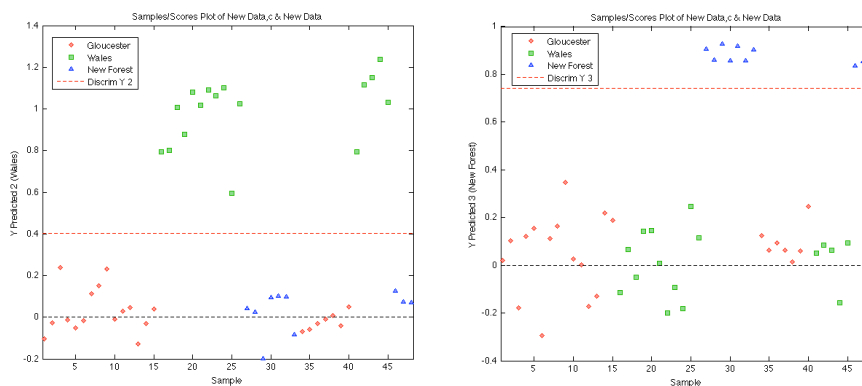
259



Predictions for Class Gloucester



Predictions for Class Wales

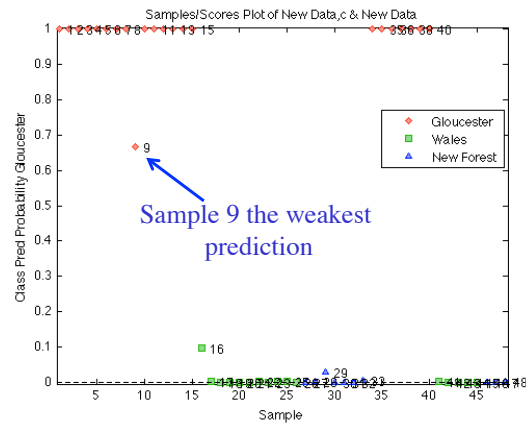


No false positives or negatives

261

EIGENVECTOR
RESEARCH INCORPORATED

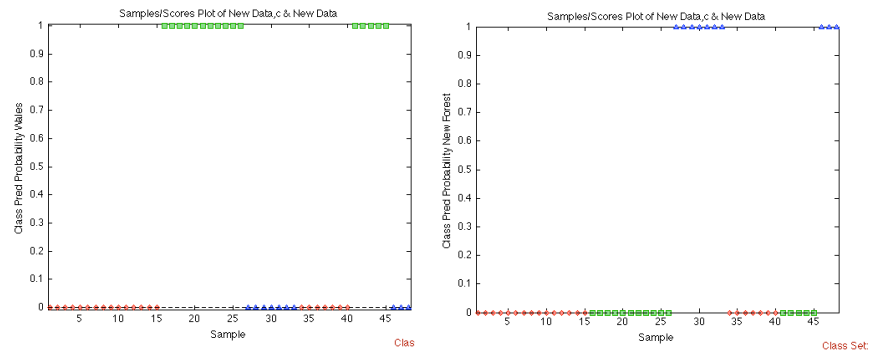
Class Prediction Probability for Gloucester



262



Class Prediction Probability for Wales and New Forest



263



[Return to Slides](#)

Dangers in using PLS-DA

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

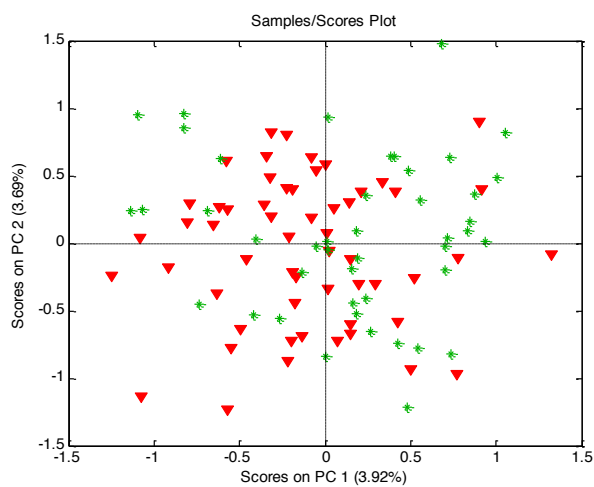
Assign samples to two classes at random –
Y-Block

Create a similar **Test Set** of 100 samples

264



PCA Scores Plot of Learning Set

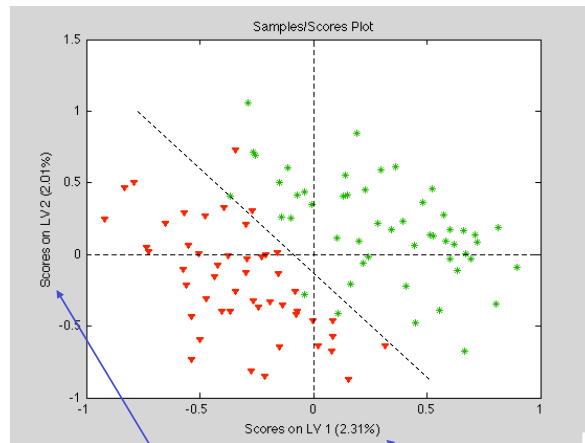


Totally Random
distribution of
samples and classes

265



PLS-DA Scores Plot of Learning Set



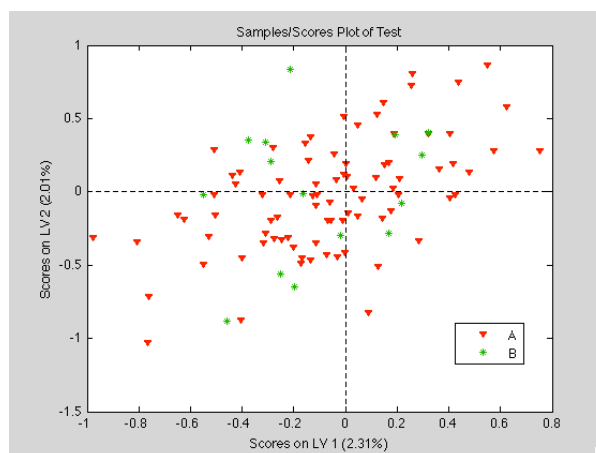
PLSDA has found differences between the two classes in the Learning Set

266

Notice small amount of variance captured.



The Test Set Projected onto the Model Scores Space



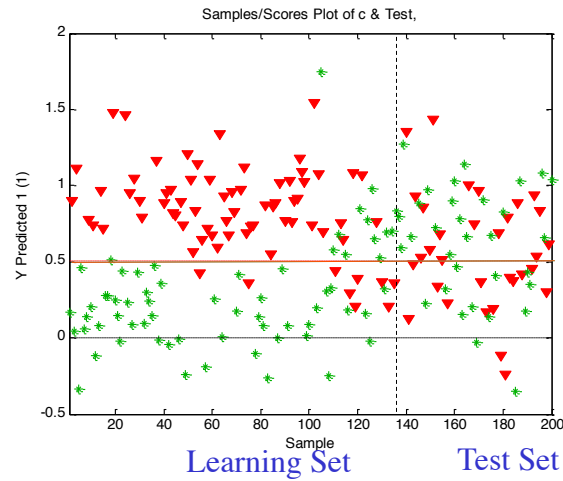
The PLSDA Model fails to distinguish the samples in the Test Set

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

267



Classification of Learning Set and Test Set



268



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

269



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.

270



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.

271

