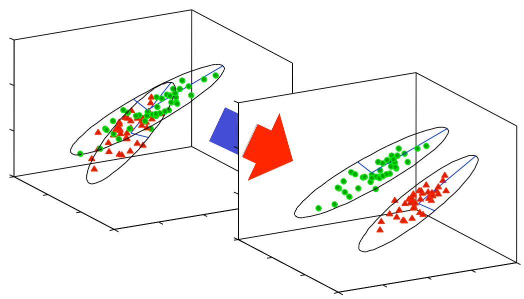
Advanced Chemometrics

Without Equations
(or hardly any)



EIGENVECTOR RESEARCH INCORPORATED

Contact Information

Eigenvector Research, Inc.
196 Hyacinth Road
Manson, WA 98831
USA

Barry M. Wise
509-662-9213
bmw@eigenvector.com

Neal B. Gallagher
(509) 687-1039
nealg@eigenvector.com

EIGENVECTOR RESEARCH INCORPORATED

Outline

- Introduction
- PCA Review
- PLS Regression Review
- Advanced Preprocessing
- Variable Selection
- Summary

3

EIGENVECTOR RESEARCH INCORPORATED

Chemometrics - Use of Mathematics, Chemistry, Physics 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 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

Chemometrics Tools

- Simple exploratory analysis (e.g., PCA and PLS) are useful and help us understand the data.
  - The goal is to see trends and gain a better understanding about the measurements and system generating the data.
  - Can provide insight into how to preprocess the data
- Mathematical tools allow us to extract information from the signal (typically multivariate) that isn't always easy to see.

5

EIGENVECTOR RESEARCH INCORPORATED

Some Resources

- Books
  - *Multivariate Calibration*, H. Martens and T. Næs, John Wiley & Sons Ltd. (1989) ISBN 0-471-90979-3
  - *Techniques and Applications of Hyperspectral Image Analysis*, Grahn, H. F.; Geladi, P., Eds. John Wiley & Sons: West Sussex, England (2007).
  - Smilde, A., Bro, R., and Geladi, P., "Multi–way Analysis with Applications in the Chemical Sciences", John Wiley & Sons, New York, NY (2004).
  - Magnus, J.R. and Neudecker, H., "Matrix Differential Calculus with Applications in Statistics and Economics, Revised Edition", John Wiley & Sons, New York, NY (1999).
- Journals
  - Journal of Chemometrics; Chemometrics and Intelligent Laboratory Systems; Analytical Chemistry; Analytica Chemica Acta; Applied Spectroscopy; Critical Reviews in Analytical Chemistry; Journal of Process Control; Computers in Chemical Engineering; Technometrics
- Special Journal Papers
  - Sanchez, E. and Kowalski, B.R., "Tensorial Calibration: II. Second Order Calibration", *J. Chemometrics*, **2**, 247–263 (1988).
  - Martens, H., Nielsen, J. P., Engelsen, S. B., "Light Scattering and Light Absorbance Separated by Extended Multiplicative Signal Correction. Application to Near–Infrared Transmission Analysis of Powder Mixtures", *Anal. Chem.*, **75**(3), 394–404 (2003).

6

EIGENVECTOR RESEARCH INCORPORATED

Advanced Chemometrics

- Advanced concepts combine our understanding of the physics and chemistry of the system, and knowledge of how the mathematical tools work to provide better experimental designs and to …
- maximize signal-to-noise ➔ signal-to-clutter

7

EIGENVECTOR RESEARCH INCORPORATED

Principal Components Analysis Review

- We'll come back to "maximize signal-to-noise ➔ signal-to-clutter"
- First let's review PCA and follow through an example
  - start software, load data, perform a PCA decomposition and define PCA terms

8

EIGENVECTOR RESEARCH INCORPORATED

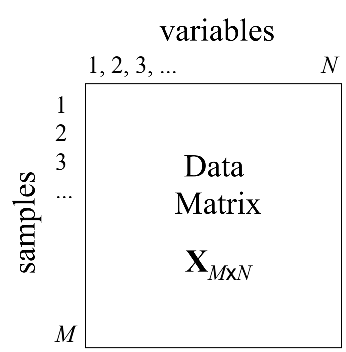
Outline

- Introduction
- PCA Review
  - Mean-centering and autoscaling
- PLS Regression Review
- Advanced Preprocessing
- Variable Selection
- Summary

9

Data Matrix **X**: Variables and Samples

- Examples of variables:
  - absorbance at each l
  - ion current at each m/e
  - pressure, temperature, flow
  - chromatographic peak area
- Examples of samples:
  - samples taken to lab
  - data samples at time points
  - data from specific batches
  - etc....



10

PCA Decomposition

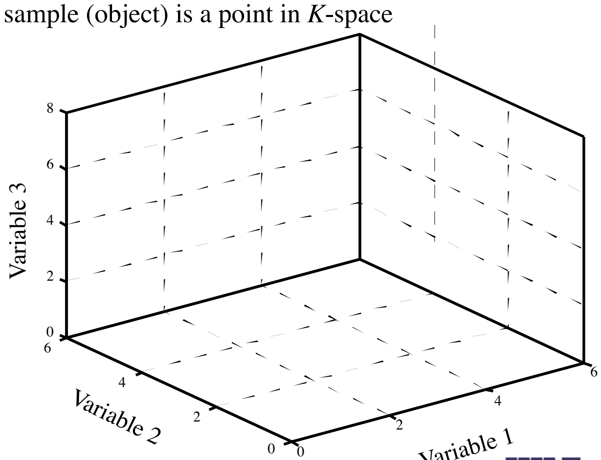
- PCA partitions a data matrix into
  - sample related information (scores) and
  - variable related information (loadings).
- Useful for multivariate exploratory data analysis.
- Scores and loadings are determined by maximizing capture of variance
  - information, sum-of-squares
    - show this graphically
  - Many methods in multivariate analysis are "factor based" – PCA factors are scores and loadings.

11

EIGENVECTOR RESEARCH INCORPORATED

Principal of Projections

- *K*-space has *K* dimensions where each variable, or measurement on an object, is a coordinate axis
- A sample (object) is a point in *K*-space

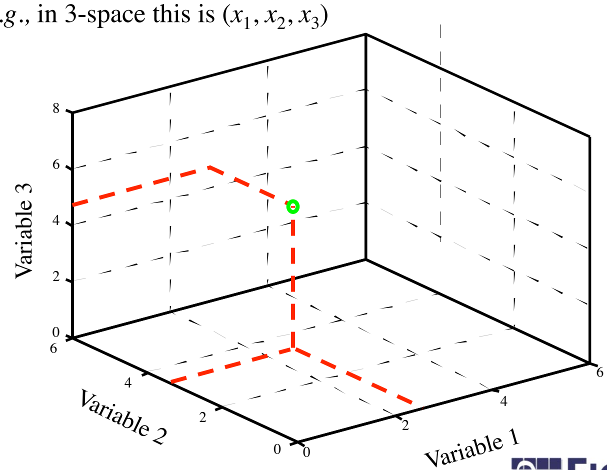


12

EIGENVECTOR RESEARCH INCORPORATED

Projection in K-Space

- The projection of an object onto the *K*-space yields the coordinates of the object in that space
- *e.g.*, in 3-space this is (x_1, x_2, x_3)

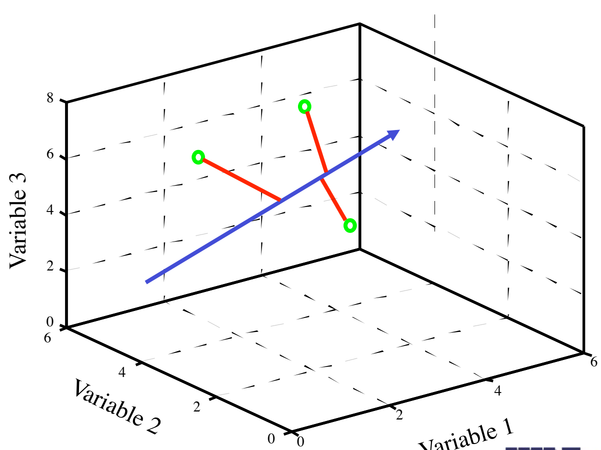


13

EIGENVECTOR RESEARCH INCORPORATED

Projection onto a Vector

- Projection lines are perpendicular to the vector

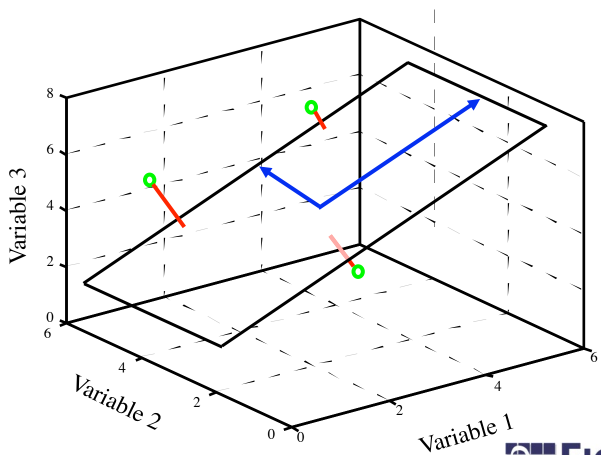


14

EIGENVECTOR RESEARCH INCORPORATED

Projection onto a Plane

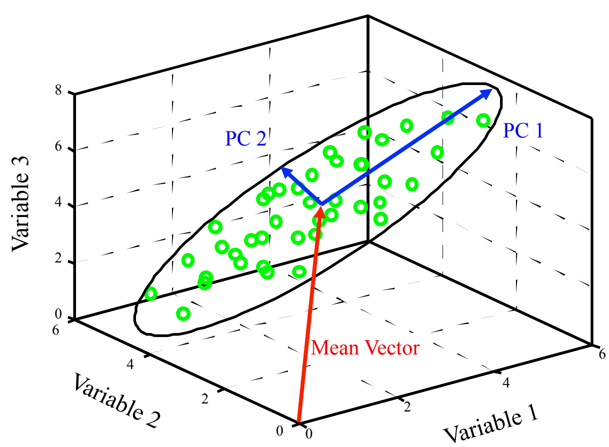
- Projection lines are perpendicular to the plane



15

EIGENVECTOR RESEARCH INCORPORATED

PCA

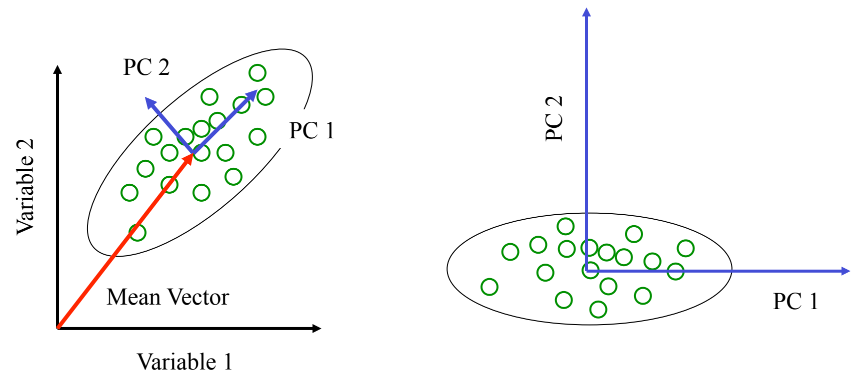


16

EIGENVECTOR RESEARCH INCORPORATED

PCA

- Geometry for 2 variables



17

EIGENVECTOR RESEARCH INCORPORATED

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 independant) to PC1 and passes through the second most variance.
- The process can be continued until the number of new PC's = number of old variables.

18

EIGENVECTOR RESEARCH INCORPORATED

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

19

EIGENVECTOR RESEARCH INCORPORATED

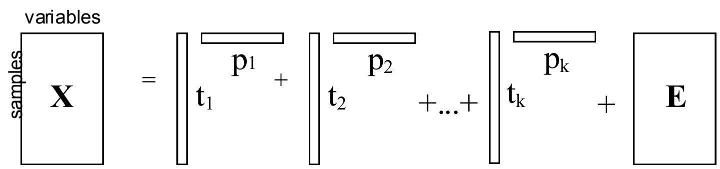
What Does PCA Give Me?

- Loadings: Compositions of the new PC axes in terms of the old variables. May be able to interpret the loadings in chemical terms, shows how variables are correlated.
  - 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

20

EIGENVECTOR RESEARCH INCORPORATED

PCA



For **X** with *M* samples and *N* variables, the PCA decomposition is:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \ldots + \mathbf{t}_K\mathbf{p}_K^T + \ldots + \mathbf{t}_R\mathbf{p}_R^T$$

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \ldots + \mathbf{t}_K\mathbf{p}_K^T + \mathbf{E} = \mathbf{T}_K\mathbf{P}_K^T + \mathbf{E}$$

$R \le \min(M,N)$ is the mathematical rank of the data.

$K \ll R$ is the pseudo- or chemical-rank of the data.

The $\mathbf{p}_i$ are eigenvectors of the covariance matrix of **X** and λ_i are eigenvalues. Amount of variance captured by each $\mathbf{t}_i\mathbf{p}_i^T$ is proportional to λ_i.

21

EIGENVECTOR RESEARCH INCORPORATED

Properties of PCA

- $\mathbf{t}_i$,$\mathbf{p}_i$ ordered by amount of *variance captured* $\propto \lambda_i$
- the chemical-rank *K* is the number of PCs that captures other than random noise
- $\mathbf{t}_i$ or *scores* form an orthogonal set $\mathbf{T}_k$ which describe relationship between *samples*
- $\mathbf{p}_i$ or *loadings* form an orthonormal set $\mathbf{P}_k$ which describe relationship between *variables*
- scores and loadings plots are interpreted in pairs
  - *e.g.* plot $\mathbf{t}_i$ vs sample number and $\mathbf{p}_i$ vs variable number
- it is useful to plot $\mathbf{t}_{i+1}$ vs. $\mathbf{t}_i$ and $\mathbf{p}_{i+1}$ vs. $\mathbf{p}_i$

22

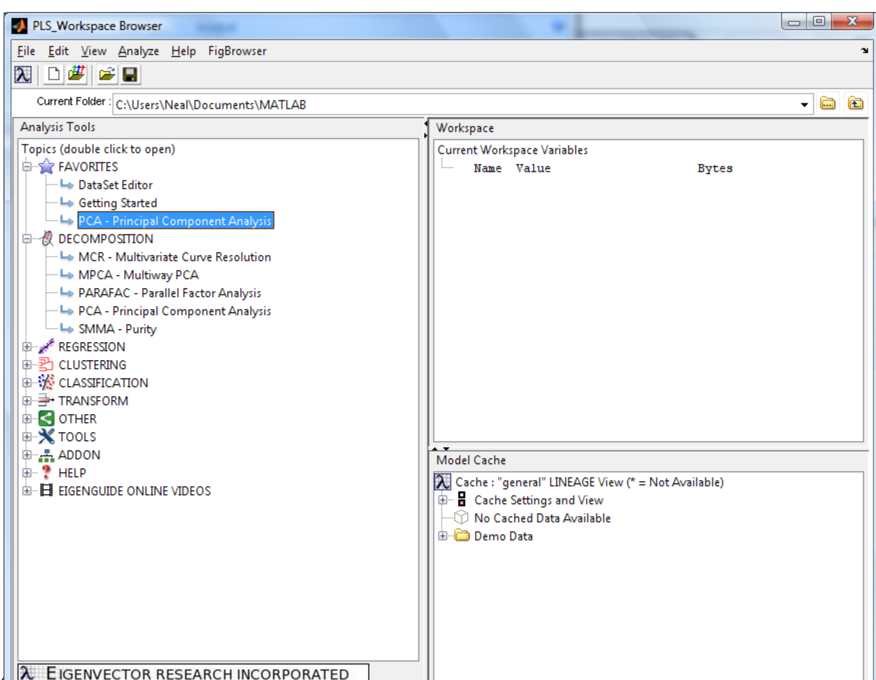
EIGENVECTOR RESEARCH INCORPORATED

Example: Olive Oils

- Use FT-IR spectra and PCA for pattern recognition to distinguish authentic olive oil from counterfeit or adulterated olive oil.
- Shall see some special properties associated with Spectral Data.
  - Dahlberg, D.B., Lee, S.M, Wegner, S.J. and Vargo, J.A., "Classification of Vegetable Oils by FT-IR," *Appl. Spec.*, **51**(8), 1118-1124 (1997).
  - FT-IR spectra (3600 - 600 cm^{-1}) using a fixed pathlength NaCl cell

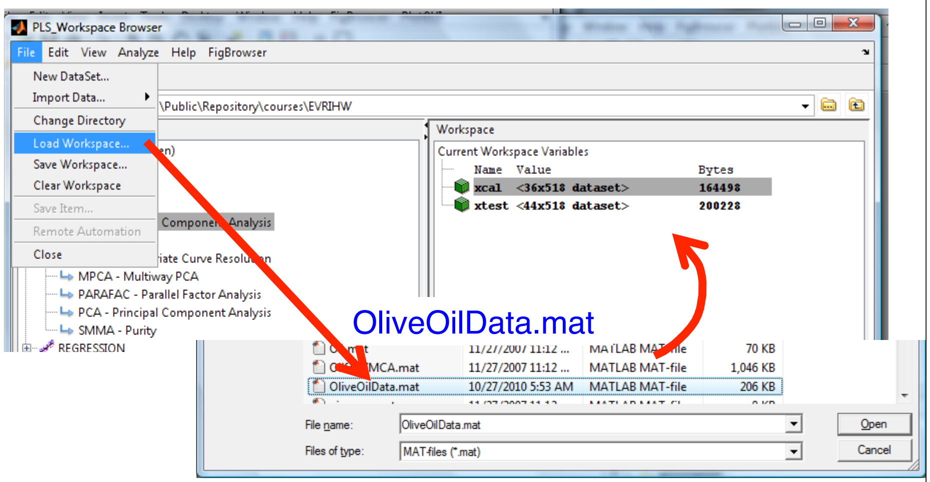
23

EIGENVECTOR RESEARCH INCORPORATED



24

Load OliveOilData.mat



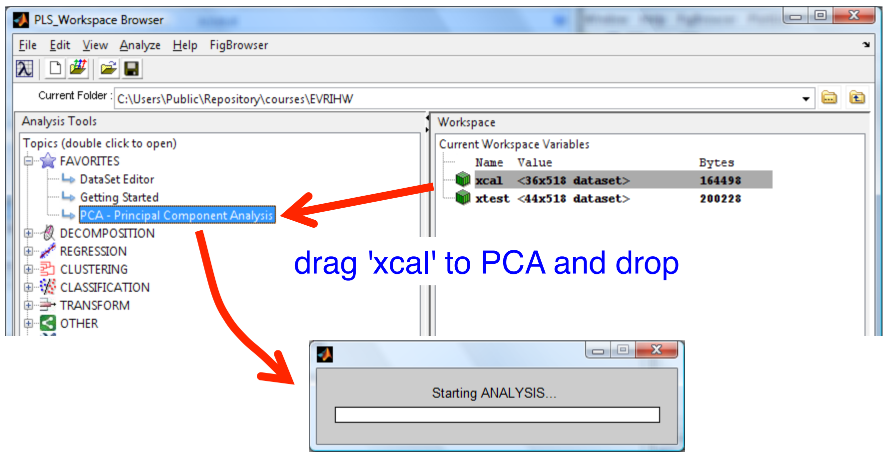
25

Olive Oil Samples

Learning set: xcal Start with this data set

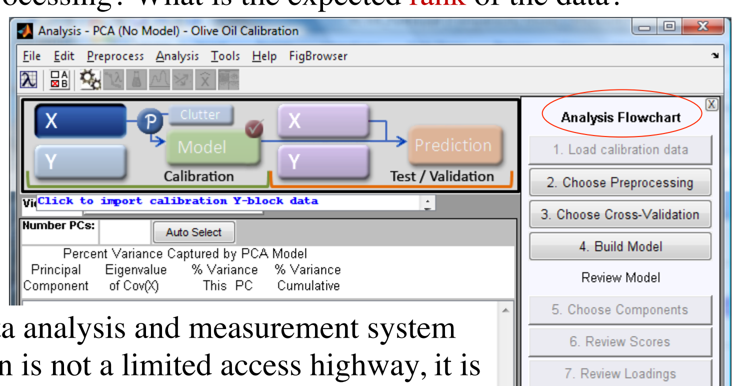
Corn Oil	9 samples	(#1-9)
Olive Oil	15 samples	(#10-24)
Safflower Oil	8 samples	(#25-32)
Corn Margarine	4 samples	(#33-36)

Start Analysis for PCA



27

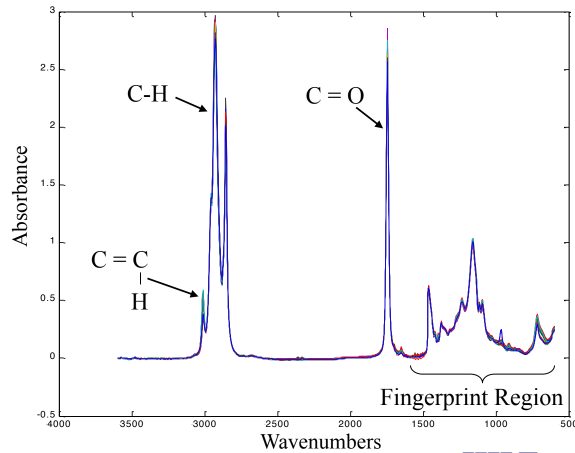
Plot the data. (for this example let's include all the variables)
Use knowledge and logic during the analysis – this is *not* a black box.
Before modeling ask, "what will PCA give me for this data and this preprocessing? What is the expected rank of the data?"



Data analysis and measurement system design is not a limited access highway, it is more like a worn path in the dirt.
Often what is learned at one step leads us back to the beginning.

28

Spectra of 36 Sample Learning Set



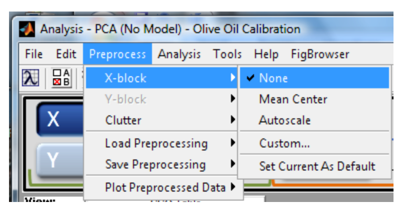
Notice how spectra look alike.

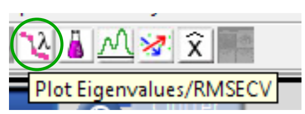
29

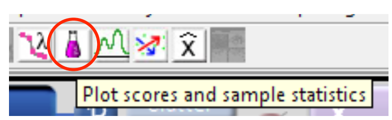
EIGENVECTOR RESEARCH INCORPORATED

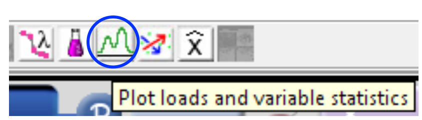
Try PCA

- Use no preprocessing
- plot the eigenvalues
  - choose number of PCs
- plot scores and loadings
  - interpret the results





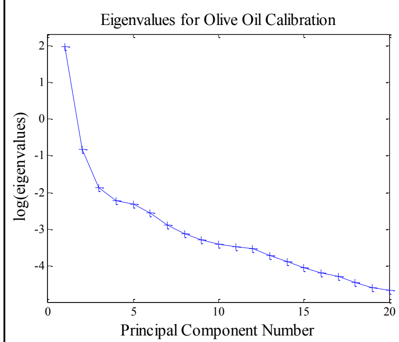




30

EIGENVECTOR RESEARCH INCORPORATED

Choosing Number of PCs

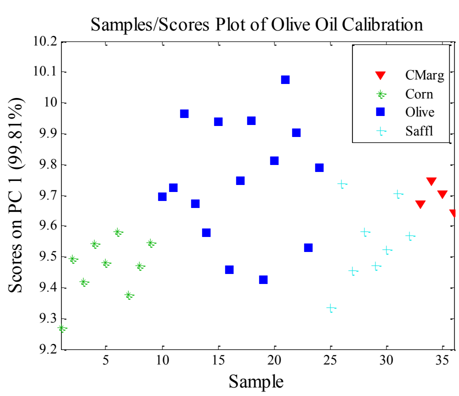


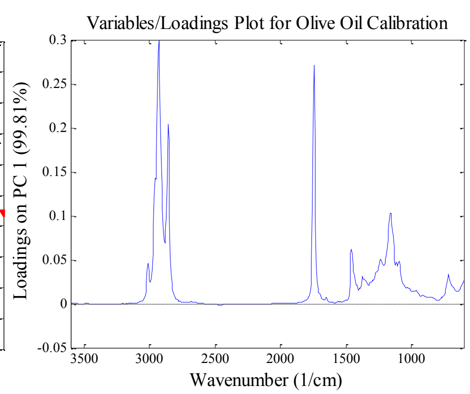
- It's not always easy
- In exploratory analysis it doesn't really matter
- Compare total % variance of model with error of data.
- Eigenvalue Plot - PCs before a break.
- PRESS plot from cross-validation - PCs at first minimum or near plateau.
- Chemical intuition to choose between conflicting results.
  - e.g., do the PCs discriminate?

31

EIGENVECTOR RESEARCH INCORPORATED

Scores and Loadings, PC 1

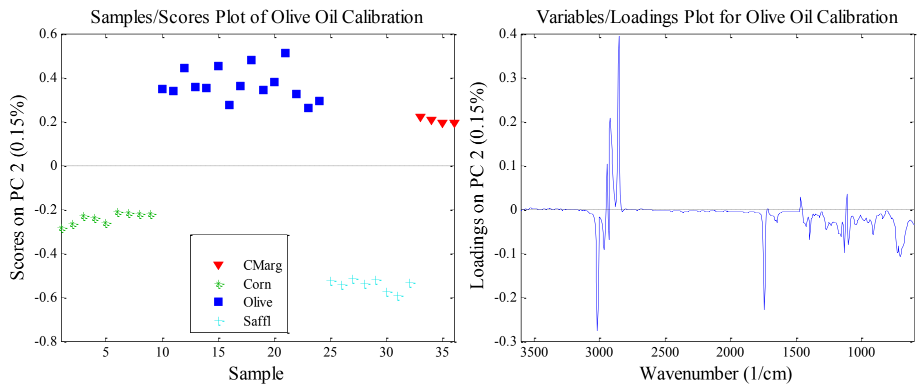




32

EIGENVECTOR RESEARCH INCORPORATED

Scores and Loadings, PC 2



33

EIGENVECTOR RESEARCH INCORPORATED

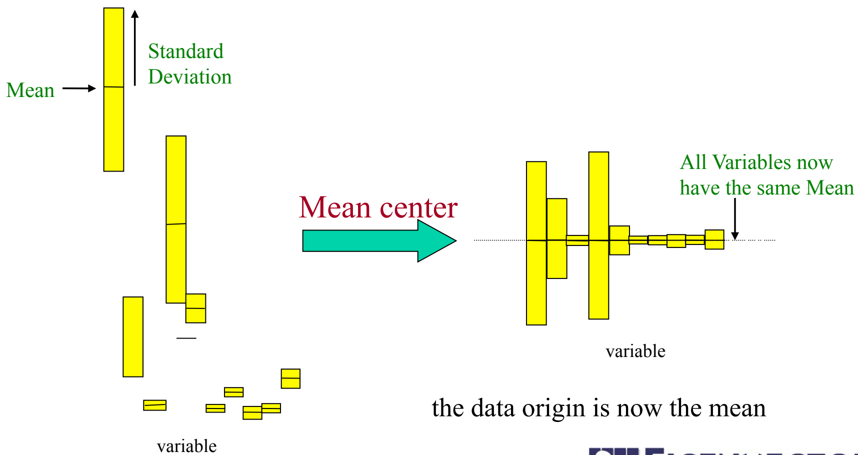
PCA Summary

- No preprocessing
  - PC 1 captured variance that was in the general direction of the mean
    - although it is not strictly the mean of the data
  - PC 2 discriminated the oils
    - some variables associated with differences between the oils were seen on PC 2
    - discrimination wasn't great, can we do better?
  - PCA is designed to capture sum-of-squares from the origin
    - that's why PC 1 was in direction of the mean!

34

EIGENVECTOR RESEARCH INCORPORATED

Repeat the PCA with Mean-Centering

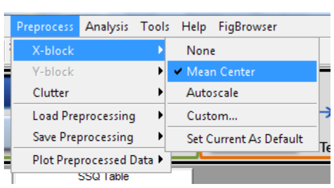


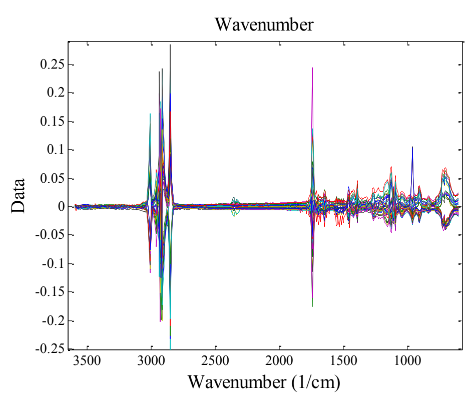
the data origin is now the mean

35

Mean-Centering

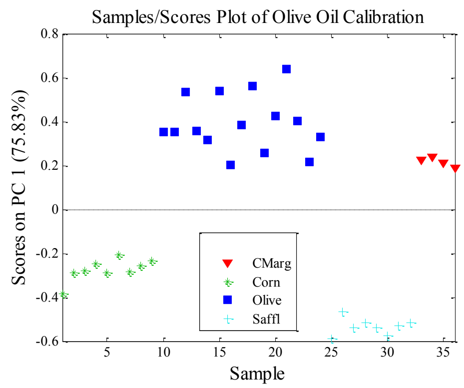
- PCA now captures sum-of-squares about the mean of the data

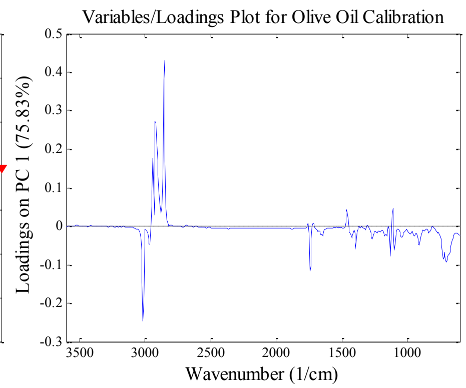




36

Scores and Loadings, PC 1





37

EIGENVECTOR RESEARCH INCORPORATED

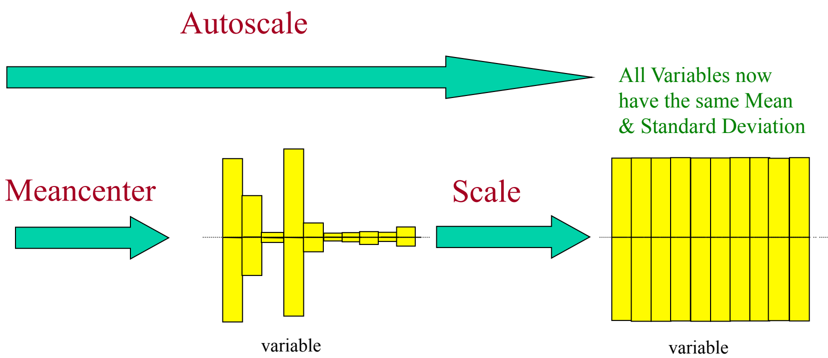
PCA Summary

- Mean-centering
  - removing the mean now focused PC 1 on variance about the mean and PC 1 discriminated the oils
    - we're bringing relevant variance closer to the top
  - median-centering can be used when there are expected to be outliers that might influence the mean
    - the outliers are easier to identify and then remove
  - additionally, we identified
    - regions with little or no signal
    - sloping baseline variability
  - can we do better, how about auto-scaling?

38

EIGENVECTOR RESEARCH INCORPORATED

Autoscaling



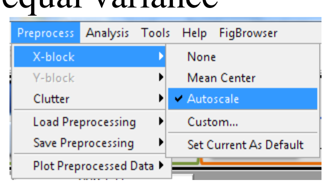
Remember: autoscaling includes mean centering

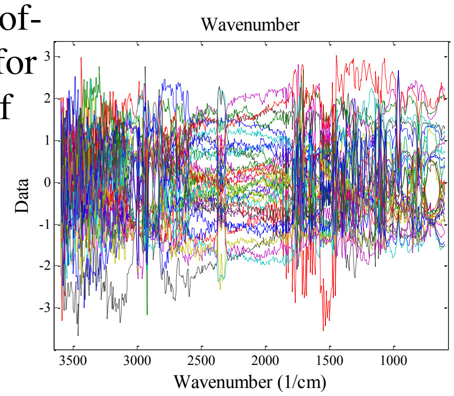
39

EIGENVECTOR RESEARCH INCORPORATED

Auto-Scaling

- PCA now captures sum-of-squares about the mean for the data with variables of equal variance

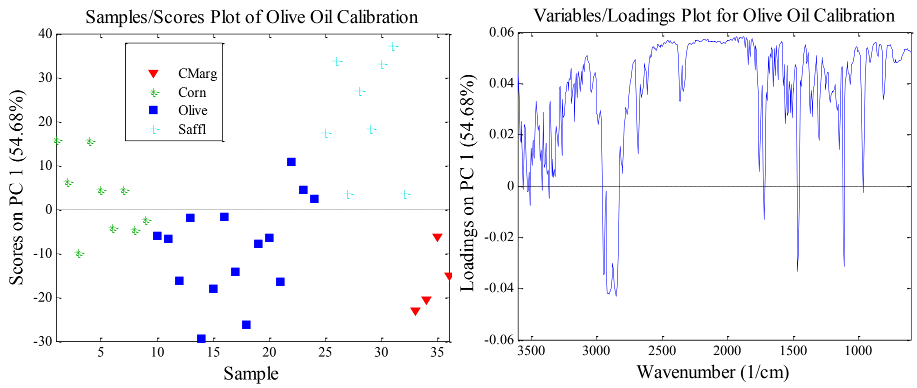




40

EIGENVECTOR RESEARCH INCORPORATED

Scores and Loadings, PC 1



41

EIGENVECTOR RESEARCH INCORPORATED

PCA Summary

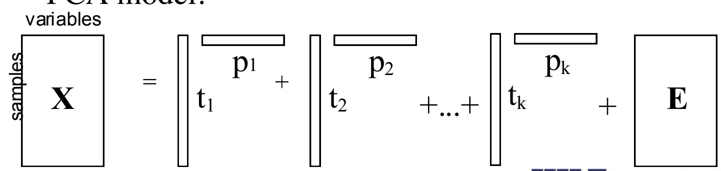
- Autoscaling
  - PC 1 no longer discriminates
    - giving all the variables an equal weight "blew up" noisy variables that had small signal and subsequently added lot's of sum-of-squares
    - we're bringing irrelevant variance closer to the top
  - for many data sets, autoscaling is a good thing, but not often used in spectra
    - autoscaling ~assumes that each variable has a similar S/N
      - but clearly not the case over the entire spectral range
    - often used when variables are of different units
      - e.g., in engineering applications

42

EIGENVECTOR RESEARCH INCORPORATED

Q Statistic in PCA

- Recall that the PCA model was truncated to keep only *K* PCs.
- What about **E**? **E** is the lack of fit.
- The Q statistic is the sum-of-squares of each row of **E** and is a measure of lack of fit of each sample.
  - It is a measure of the distance from the plane of the PCA model.



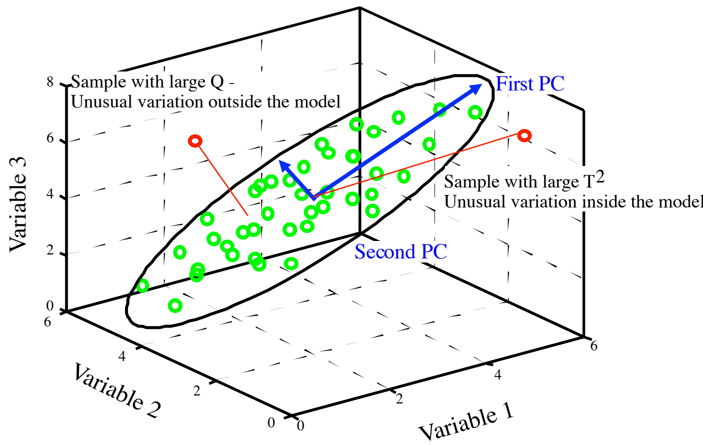
43

Hotelling's T^2

- Hotelling's T^2 statistic can be calculated from the PCA scores.
- T^2 accounts for the different amounts of variance in each direction to calculate a distance from the origin within the plane of the PCA model.

44

Geometry of Q and T^2



45

EIGENVECTOR RESEARCH INCORPORATED

Outline

- Introduction
- PCA Review
- PLS Regression Review
  - cross-validation
  - Savitsky-Golay
  - model validation
- Advanced Preprocessing
- Variable Selection
- Summary

46

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 variables are measured that are available in a timely manner to predict something that is more difficult (or more expensive) to obtain.

*"You Can't Always Get What You Want," Rolling Stones, Let it Bleed (1969)

47 EIGENVECTOR RESEARCH INCORPORATED

Regression

X-Block $\xrightarrow{\text{Regression Model}}$ **Y**-Block

What's measured — What's desired

PCA was used to explore the correlation structure within a single data block **X**.

Regression analysis identifies the dependency between two blocks of data **X** and **Y**.

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

48 EIGENVECTOR RESEARCH INCORPORATED

Many Forms of Regression

- Classical Least Squares (CLS)
  - Generalized least squares
  - Extended least squares
- Multiple Linear Regression (MLR)
- Principal Components Regression (PCR)
- Partial Least Squares Regression (PLS)

$\mathbf{Xb} = \mathbf{y} + \mathbf{e}$ (this is our focus)

$\mathbf{XB} = \mathbf{Y}+\mathbf{E}$ (PLS-2: multivariate $\mathbf{Y}$)

49

EIGENVECTOR RESEARCH INCORPORATED

PLS Description

- PCA decomposes **X** into factors called PCs
- PLS decomposes **X** (and **Y**) into latent variables
- Selection of the number of LV's is ~more important in PLS than in PCA but it's also a bit easier
  - Cross-validation

50

EIGENVECTOR RESEARCH INCORPORATED

Cross-Validation

- Divide data set into *J* sample subsets to leave out one at a time.
- For ***each subset***:
  - build a PLS model using all samples in the ***remaining*** subsets (i.e., build *J* models) and using different numbers of LVs (1,2,…)
  - apply the model to predict the J^{th} subset samples
  - calculate PRESS (Predictive Residual Sum of Squares) for the subset samples and sum over all *J* subsets and LVs:

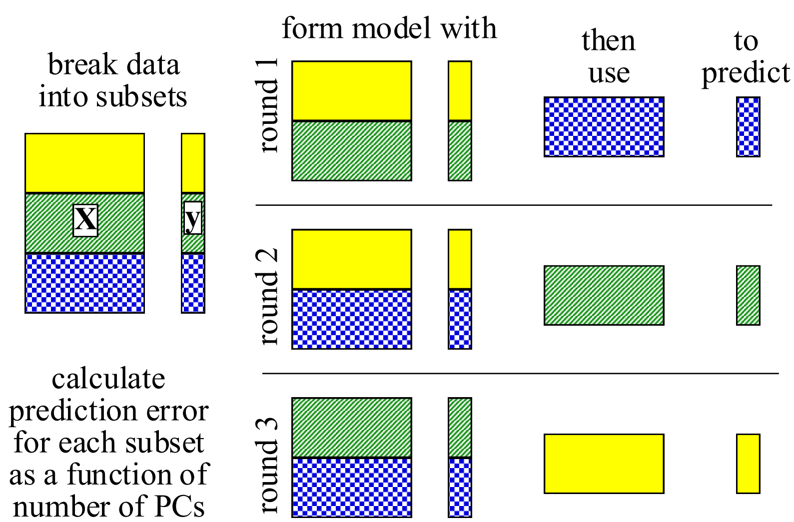
$$\mathbf{e}^2 = \left(\mathbf{y} - \mathbf{X}\mathbf{b}\right)^2$$

- Look for minimum or "knee" in PRESS curve

51

EIGENVECTOR RESEARCH INCORPORATED

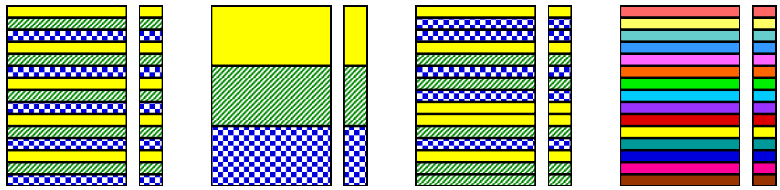
Cross-validation Graphically



52

EIGENVECTOR RESEARCH INCORPORATED

Formation of Test Sets



"Venetian blinds" - OK when data already in random order	contiguous blocks-best for time series	random selection- usually repeated several times	leave-one-out, used when not much data available

What else? **Custom selection, based on prior knowledge!**

53

EIGENVECTOR RESEARCH INCORPORATED

Cross-validation Considerations

- Cross-validation method selection criteria
  - *Number* of objects in dataset
  - *Order* of objects in dataset
  - *Objective* of cross-validation (specific type of error?)
  - Presence/absence of *replicates*
- "Traps" to avoid
  - "Repeat sample trap"
    - Repeat measurements in both model and test set
  - "External subset selection trap"
    - Test set "space" outside of model set "space"

54

EIGENVECTOR RESEARCH INCORPORATED

Cross-Validation Rules of Thumb

- Divide data set into ~square root of number of samples subsets.
- "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)

55

EIGENVECTOR RESEARCH INCORPORATED

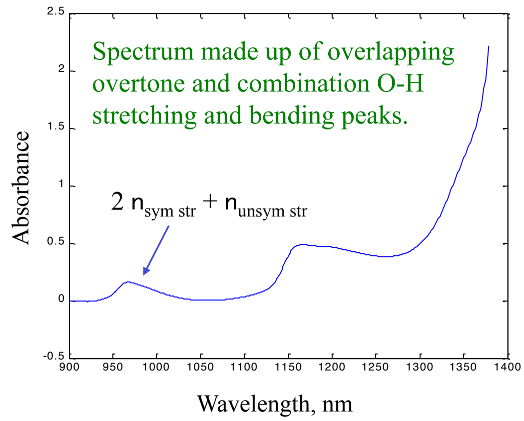
Example Application of PLS

- Estimate the concentration of NaOH in aqueous caustic brine solutions using SW-NIR
  - measured 12 solutions of NaCl and NaOH in water
  - peaks shift with changes in NaCl , NaOH and temperature, T
- Since T will vary in the application, T variation must be included in the Learning Set
  - although T need not be known to calibrate for NaOH, it must vary in the Learning Set for the model to be robust to changes in T

56

EIGENVECTOR RESEARCH INCORPORATED

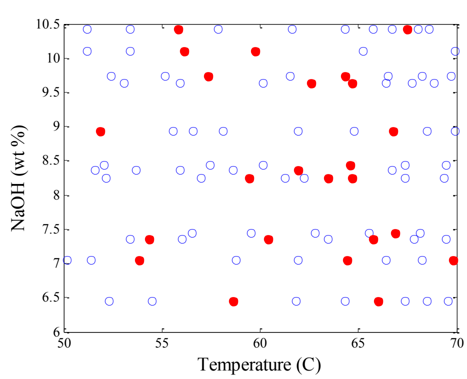
Typical SW-NIR Spectrum of Caustic Brine Solution



57

EIGENVECTOR RESEARCH INCORPORATED

Learning and Test Sets Randomly Selected Samples



○ Learning Set 71 samples

● Test Set remaining 24 samples

T: 50 to 70 C

58

EIGENVECTOR RESEARCH INCORPORATED

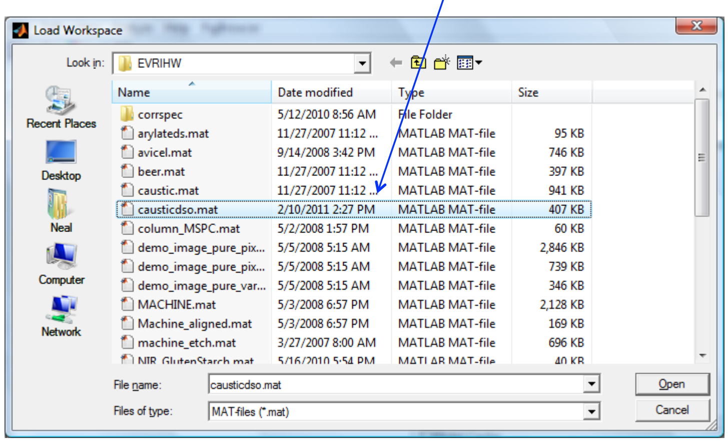
PLS Example

- How to preprocess?
  - just mean-center for now
- How many latent variables?
  - cross-validation using venetian blinds
  - split the data sqrt(71)~8 times
  - examine out to 20 LVs (expect true number < 20)

59

EIGENVECTOR RESEARCH INCORPORATED

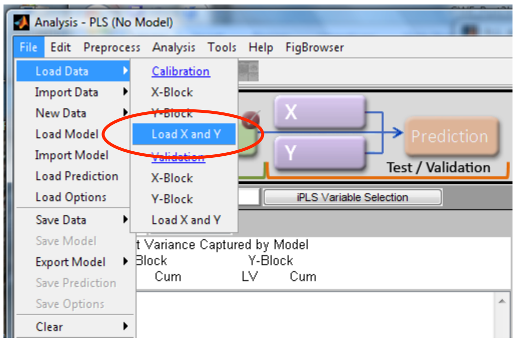
- Analysis: File: Clear: All
- Browser: File: Clear Workspace
- Browser: File: Load Workspace: causticdso.mat

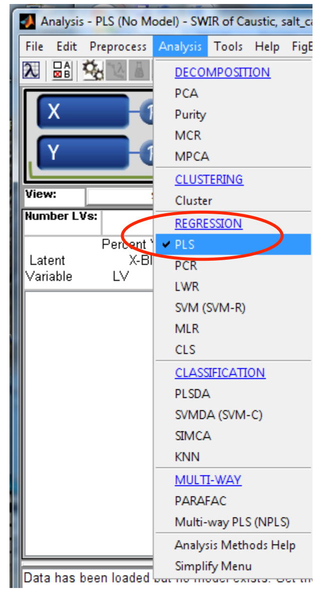


60

EIGENVECTOR RESEARCH INCORPORATED

- Analysis: File: Load X and Y
  - load xcal and ycal
- Analysis: Analysis: PLS

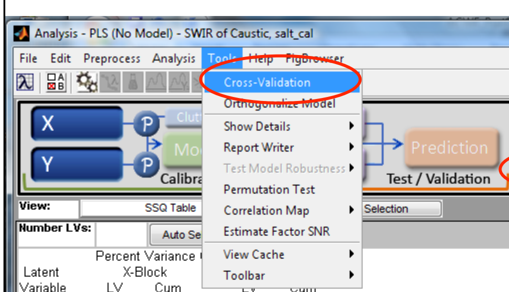


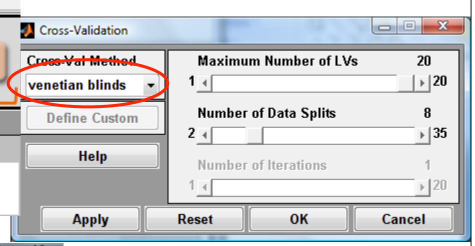


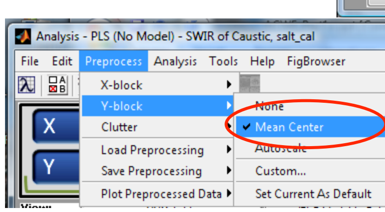
EIGENVECTOR RESEARCH INCORPORATED

61

- Analysis: Tools: Cross-Validation
- Cross-Validation: venetian blinds: OK
- Analysis: Preprocess: X,Y-Block: Mean Center



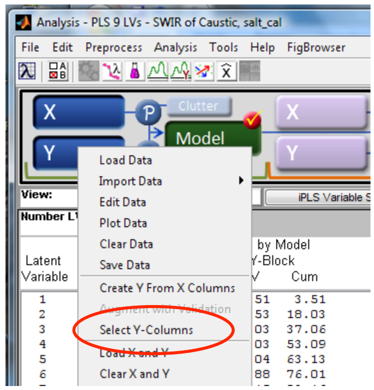


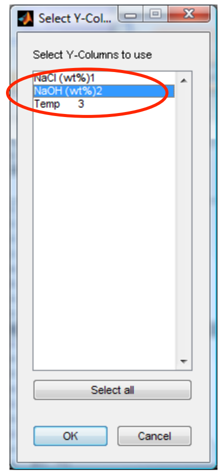


EIGENVECTOR RESEARCH INCORPORATED

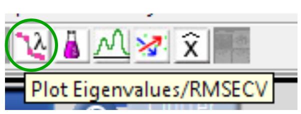
62

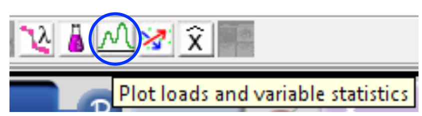
- Analysis: right-click Y: Select Y-Columns
- Select NaOH (wt%)2: OK
- Analysis: Model

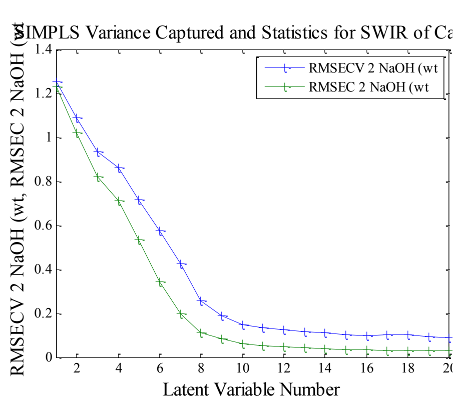


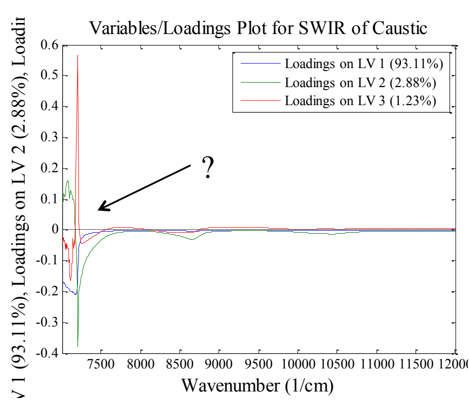


EIGENVECTOR RESEARCH INCORPORATED



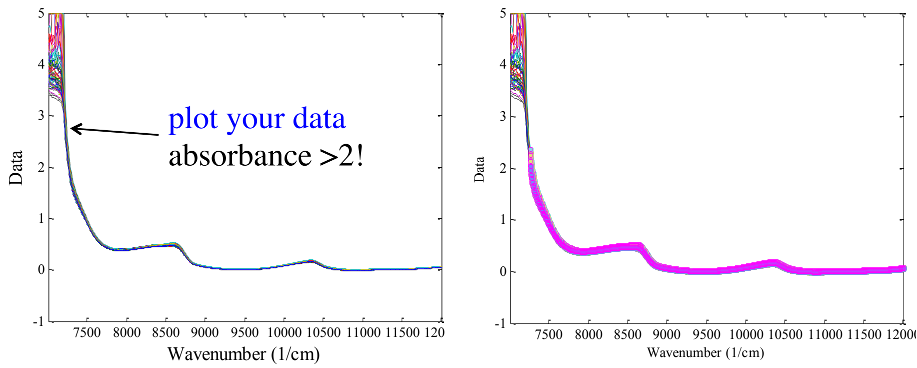






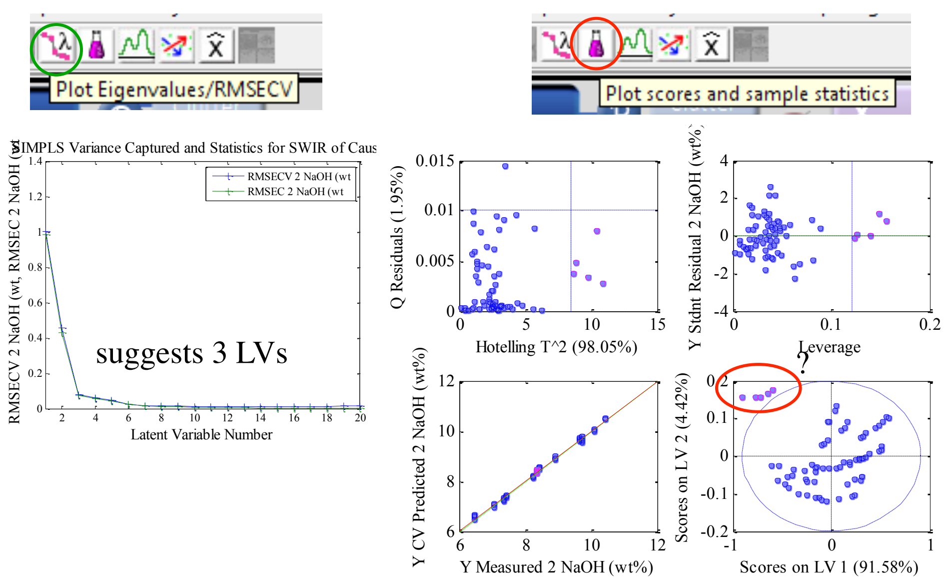
EIGENVECTOR RESEARCH INCORPORATED

- Analysis: right-click X: Plot Data
- Plot Controls: Y: Data
- Plot Controls: Select {choose values < 2}
- Plot Controls: Edit: Include Only Selection



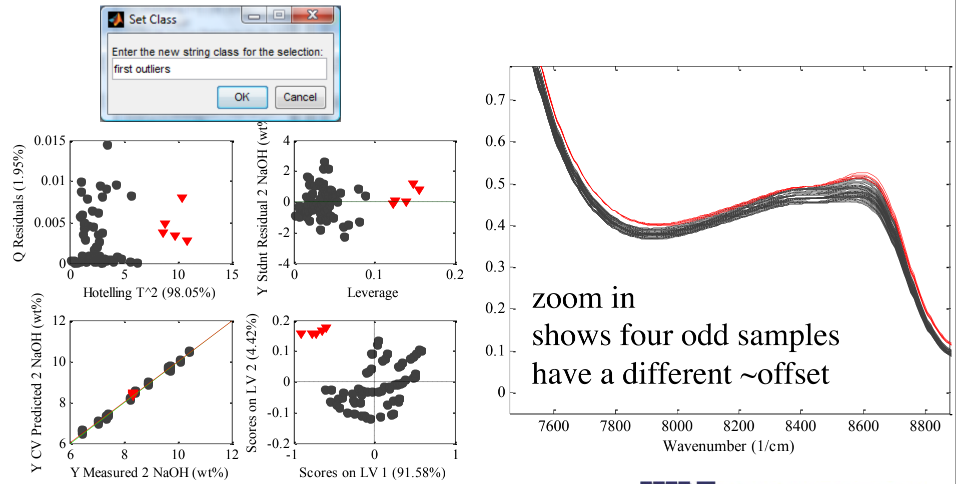
Variable selection starts by using what's known about the physics and chemistry of the measurement system.

EIGENVECTOR RESEARCH INCORPORATED



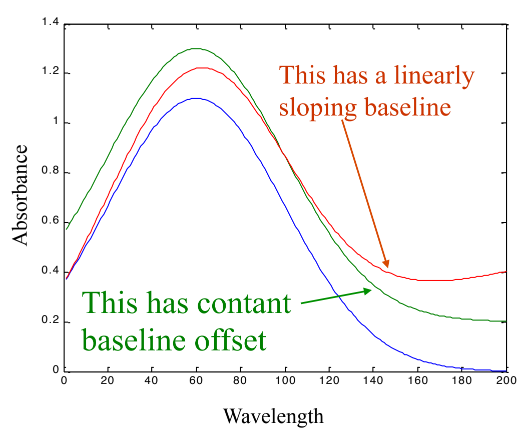
EIGENVECTOR RESEARCH INCORPORATED

- Plot Controls: Select {select the four odd samples}
- Plot Control: Edit: Set Class of Selected: "1 outliers": OK
- Plot Controls: data {select several samples}



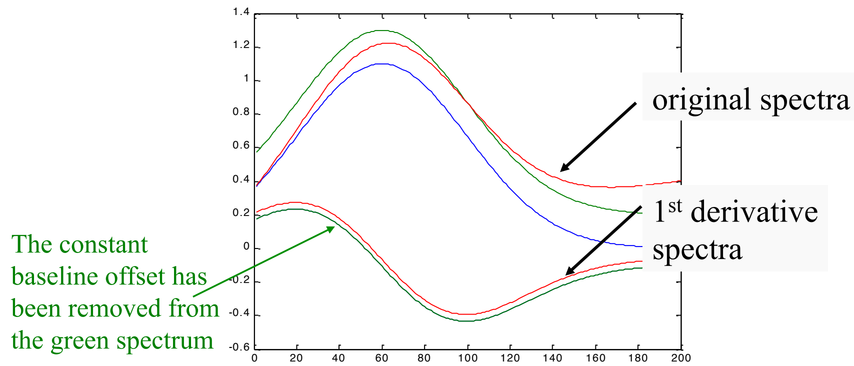
Baseline Problems Example

three identical spectra, except:



Using Derivative Spectra

Take 1st derivative of the Three Spectra

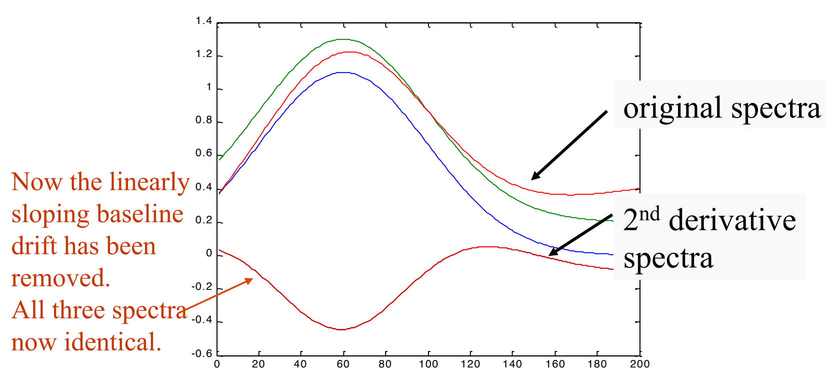


69

EIGENVECTOR RESEARCH INCORPORATED

Using Derivative Spectra

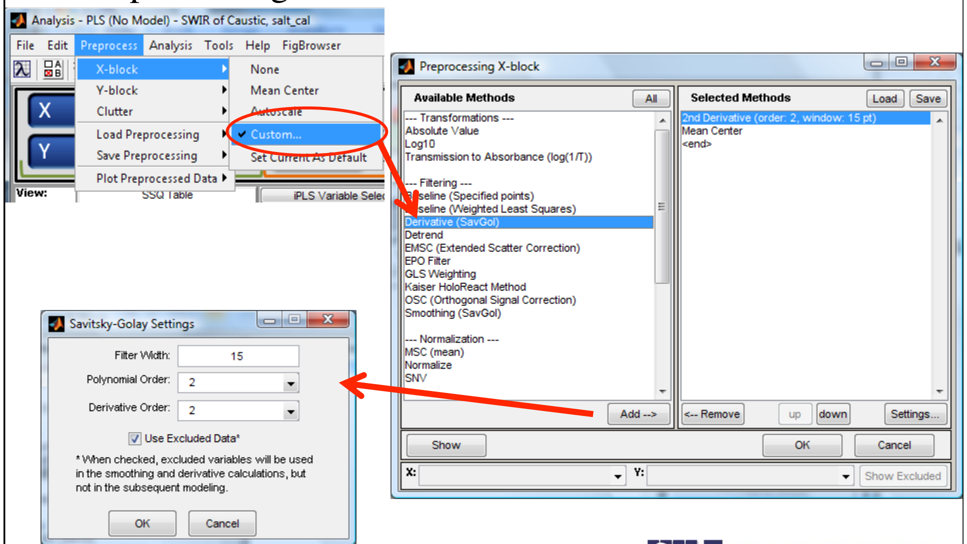
Take 2nd derivative of the Three Spectra



70

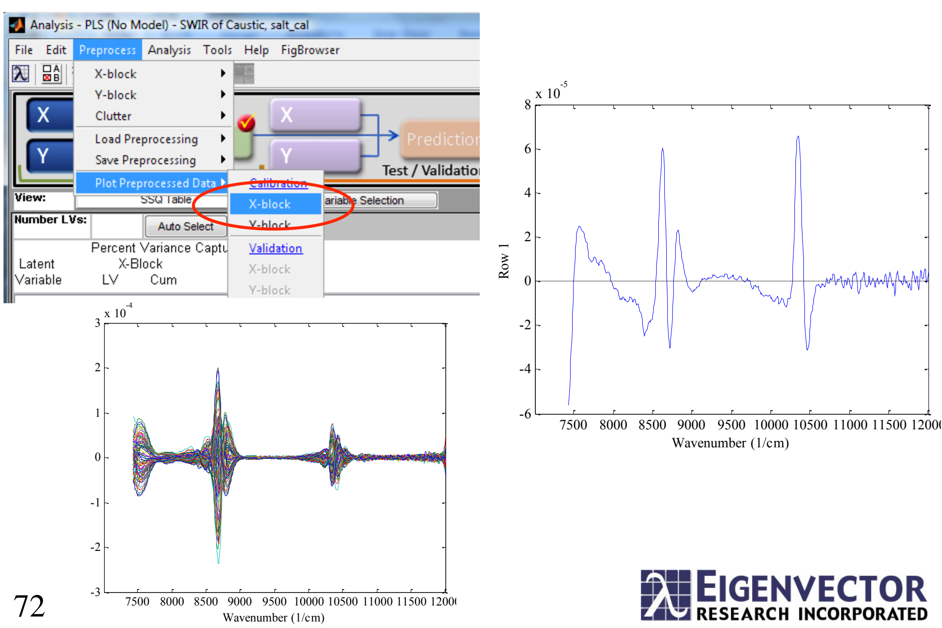
EIGENVECTOR RESEARCH INCORPORATED

- Analysis: Preprocess: X-block: Custom…
- Preprocessing X-block: Derivative: Add

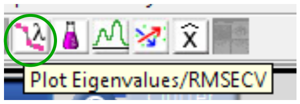


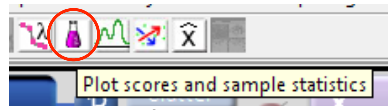
71

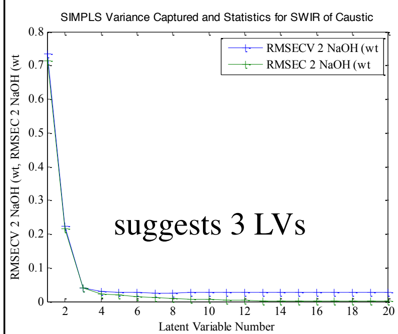
- Analysis: Preprocess: Plot Preprocessed Data: X-block

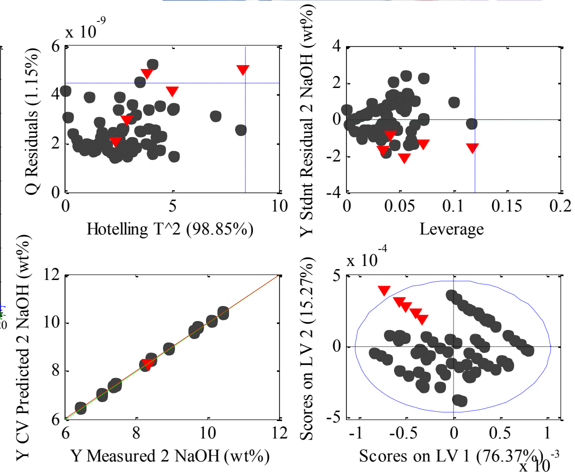


72









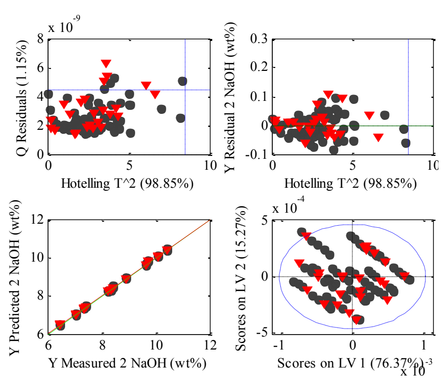
73

- Analysis: Load X, and Y-block test data / validation data and click Model

```
Linear regression model using
Partial Least Squares calculated with the SIMPLS algorithm
Developed 11-Feb-2011 09:42:59.84
Author: Neal@NEAL-VAIO_07
X-block: SWIR of Caustic, Test Set  24 by 593 Included: [1-24 ]  [ 57-649 ]
Included (in axis units): [ n/a ]  [ 7436.1-12002 ]
Preprocessing: 2nd Derivative (order: 2, window: 15 pt), Mean Center
Y-block: salt_cal  24 by 1
Included: [ 1-24 ]  [ 2 ]
Preprocessing: Mean Center
Num. LVs: 3
Cross validation: venetian blinds w/ 8 splits
RMSEC: 0.0387602
RMSECV: 0.0410341
RMSEP: 0.0410703
Bias:     0
CV Bias: -0.00018375
Pred Bias:0.00680956
R^2 Cal: 0.999081
R^2 CV: 0.99897
R^2 Pred: 0.998991
```

Percent Variance Captured by Regression Model

	-----X-Block-----		-----Y-Block-----	
Comp	This	Total	This	Total
1	76.37	76.37	68.95	68.95
2	15.27	91.64	28.20	97.14
3	7.22	98.85	2.76	99.91



74

PLS Example Summary

- Variable Selection
  - use what you know to remove irrelevant variables
  - plot your data
- Preprocessing
  - mean-centering was used to remove overall offsets
    - not mean-centering is a force fit through zero
  - Savitzky-Golay smoothing and derivatives were used to remove offsets and slopes in the spectra
- Fit and Prediction *are not* the same thing
  - model validation is very important and continues…

75

EIGENVECTOR RESEARCH INCORPORATED

Before Applying Models to Real Unknowns

Validate Them Thoroughly With a Well Designed Test Set!!

Models Do Not Last Forever

Revalidate Them Often and Rebuild Them If Necessary.

76

EIGENVECTOR RESEARCH INCORPORATED

Outline

- Introduction
- PCA Review
- PLS Regression Review
- Advanced Preprocessing
  - clutter
  - GLS, MSC, EMSC, SNV, normalization
- Variable Selection
- Summary

77

EIGENVECTOR RESEARCH INCORPORATED

What is Clutter?

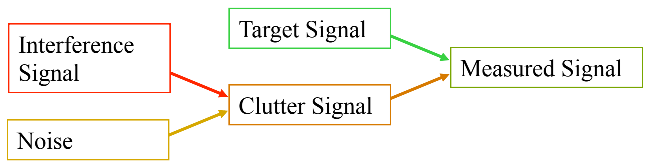
- Signal is defined as the measurement associated with the target of interest.
  - e.g., it is the part of the FTIR spectrum corresponding to discriminating the olive oils, or
  - the relationship between temperatures in a distillation column and the tray compositions
- Clutter is everything else in the measurement
  - interferences
  - noise

78

EIGENVECTOR RESEARCH INCORPORATED

Measured Signal

- Measured signal includes target
- and Clutter (X-, Y-block, …)



$$\left[\mathbf{X}_{\text{target}} + \mathbf{X}_{\text{clutter}} \pm \delta\mathbf{X}\right]\mathbf{b} = \left[\mathbf{y}_{\text{target}} + \mathbf{y}_{\text{clutter}} \pm \delta\mathbf{y}\right] + \mathbf{e}$$

- Use physics to create a linear relationship
  - non-linearity w/in X-block adds factors (digs deeper into noise)
  - non-linearity between X- and Y-blocks adds error

79

EIGENVECTOR RESEARCH INCORPORATED

Sources of Clutter

- Instrument physics
  - offset and gain changes, drift, hardware changes, smile, wavelength registration, temperature, humidity, operator …
- Sample / sampling
  - interferences chemical and physical
    - presence of other analytes
    - pathlength changes, particulate and size distribution changes, …

80

EIGENVECTOR RESEARCH INCORPORATED

Preprocessing Objective

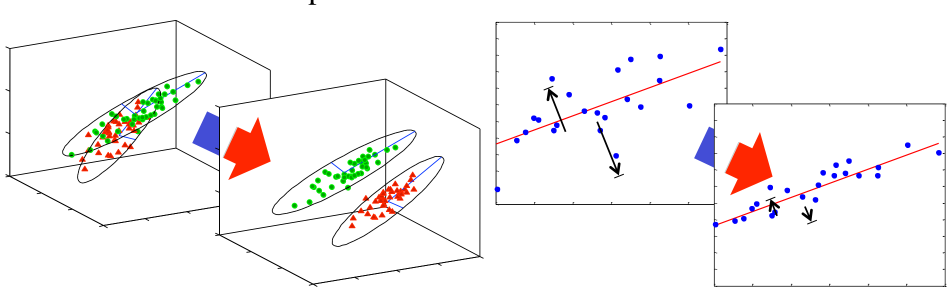
- Typical analysis methods of interest are based on maximizing capture of sum-of-squares or minimizing least-squares.
- The objective of preprocessing is to minimize variance due to clutter so that the analysis can focus on signal of interest
  - Clutter: sensor noise and the confounding effects of interferences
  - Radar Clutter Definition: (DOD, NATO) Unwanted signals, echoes, or images on the face of the display tube, which interfere with observation of desired signals.

81

EIGENVECTOR RESEARCH INCORPORATED

Advance Preprocessing

- Introduce concepts and methodologies to maximize signal-to-clutter for use in PCA and PLS
  - maximize between-class distance / within-class distance
  - minimize the prediction error



82

EIGENVECTOR RESEARCH INCORPORATED

Advanced Chemometrics

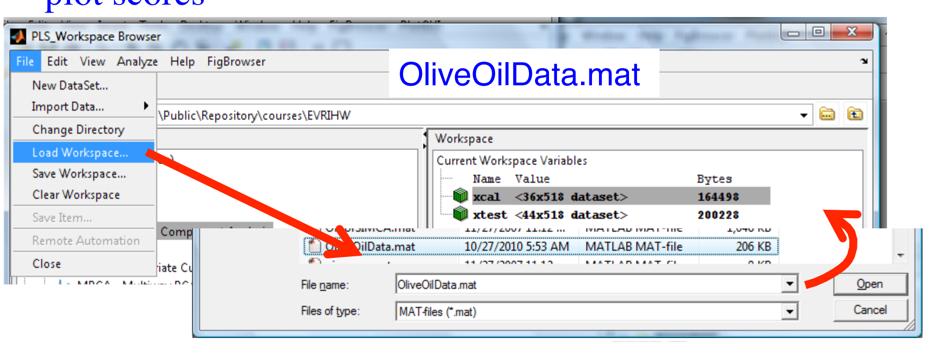
- Advanced concepts combine our understanding of the physics and chemistry of the system, and knowledge of how the mathematical tools work to provide better experimental designs and to …
- maximize signal-to-noise ➔ signal-to-clutter
- Data analysis and preprocessing should *not* be treated as a black box

83

EIGENVECTOR RESEARCH INCORPORATED

Reload OliveOilData.mat

- Analysis: File: clear all
- Browser: Clear Workspace
- Analysis, PCA, mean-centering, cross-val none, and plot scores

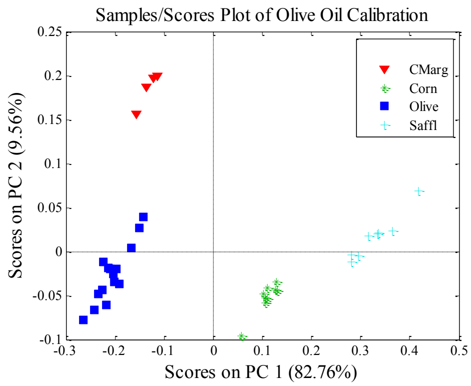


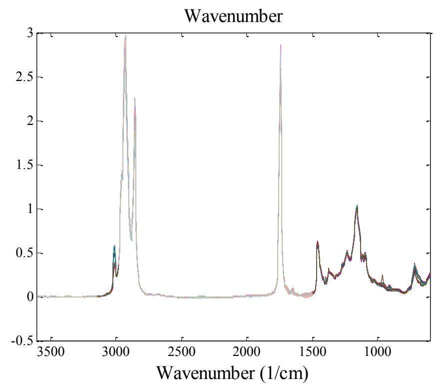
84

EIGENVECTOR RESEARCH INCORPORATED

This scores plot shows better discrimination than our previous PCA with the same preprocessing, why?

Plot your data
Plot Controls: View: Excluded Data
Variable selection removed 'irrelevant variables.'



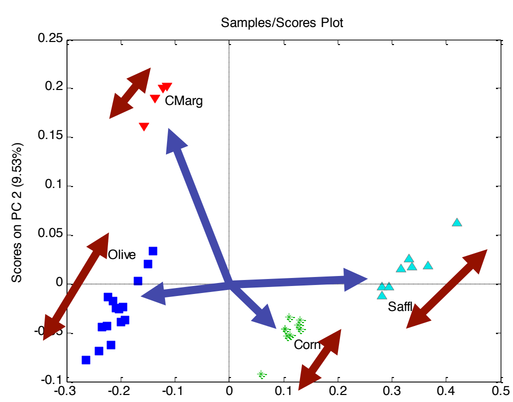


… but can we do better?

85

PCA Results

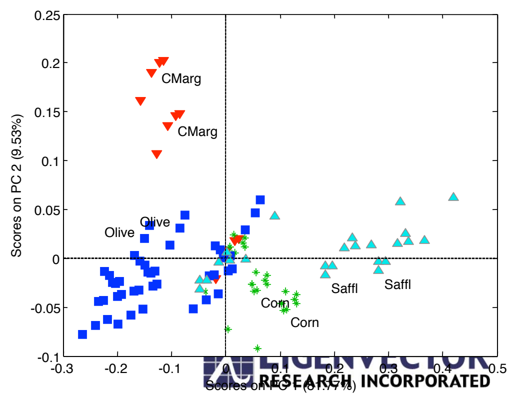
- PCA shows that the four classes in the calibration data set are separate from each other (high between class variance) but …
- have significant within class variance

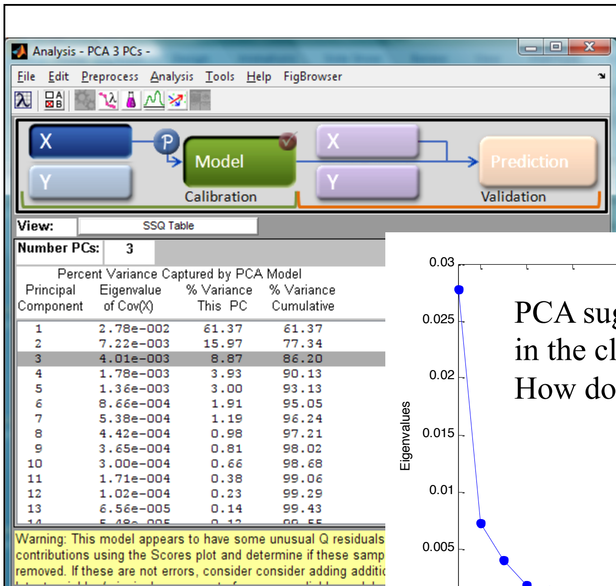


86

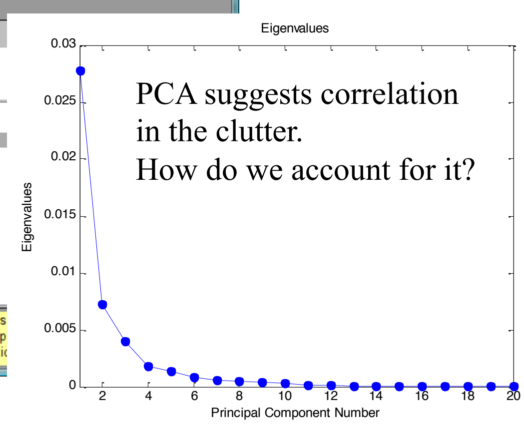
Replicates

- Ideally, replicates would lie on top of each other.
- Variance within each class is clutter variance.
  - Is it random noise? Is the clutter correlated?
- Center each class to it's own mean and do PCA on the result.





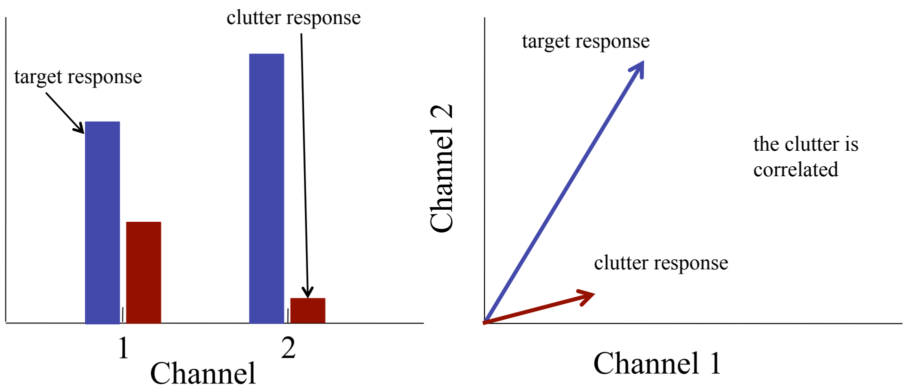
Principal Component	Eigenvalue of Cov(X)	% Variance This PC	% Variance Cumulative
1	2.78e-002	61.37	61.37
2	7.22e-003	15.97	77.34
3	4.01e-003	8.87	86.20
4	1.78e-003	3.93	90.13
5	1.36e-003	3.00	93.13
6	8.66e-004	1.91	95.05
7	5.38e-004	1.19	96.24
8	4.42e-004	0.98	97.21
9	3.65e-004	0.81	98.02
10	3.00e-004	0.66	98.68
11	1.71e-004	0.38	99.06
12	1.02e-004	0.23	99.29
13	6.56e-005	0.14	99.43



PCA suggests correlation in the clutter.
How do we account for it?

How does the clutter affect the measurements?

- Imagine a 2-channel spectrometer



89

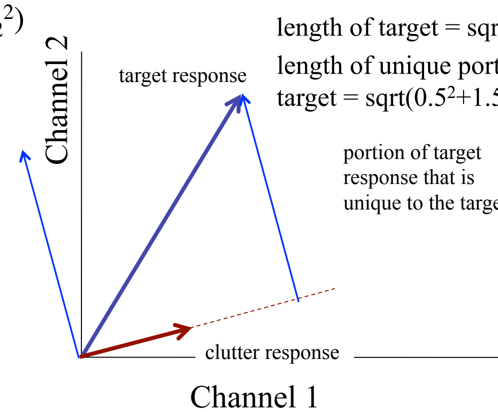
EIGENVECTOR RESEARCH INCORPORATED

How does the clutter affect the measurements?

- characterize the signal as the length of the vector $\mathrm{sqrt}(x_1^2+x_2^2)$

length of target = $\mathrm{sqrt}(1^2+2^2)=2.24$

length of unique portion of the target = $\mathrm{sqrt}(0.5^2+1.5^2)=1.58$



90

EIGENVECTOR RESEARCH INCORPORATED

Why is clutter bad?

- The signal-to-clutter is ~proportional to the length of the unique portion of the target's response.
  - in absence of clutter it was 2.24
  - in the presence of clutter it was 1.58
- In regression, clutter-to-signal is related to the estimation error.
  - higher clutter-to-signal ➔ higher estimation error
  - in the presence of clutter the estimation error is 2.24/1.58 times the error when clutter is absent

91

EIGENVECTOR RESEARCH INCORPORATED

Effect of Clutter

- The effect of clutter is to remove target signal
  - for olive oil example the target signal is the differences between the classes
- Instrument related clutter can be minimized by
  - good instrument design that accounts for the environment (noise+interferences) in which measurements are to be made
  - instrument standardization
    - remove drifts in offsets and gains that adds to the clutter
- Can't always be eliminated ➔ what to do?

92

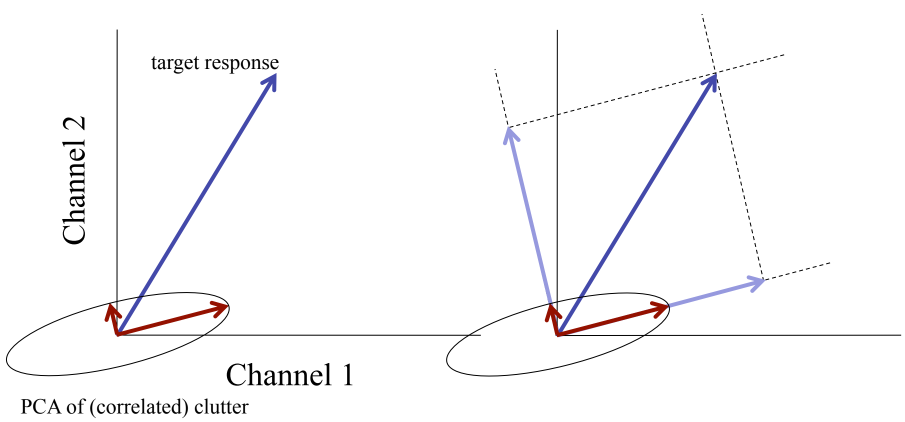
EIGENVECTOR RESEARCH INCORPORATED

Accounting for Clutter

- One method used to account for clutter is a weighting scheme
  - similar to that used in generalized least squares (GLS)
- Autoscaling scales each variable to unit variance
- GLS weighting scales each clutter direction (as determined using PCA) to unit variance
  - directions of high clutter are deweighted
  - directions of low clutter are given more opportunity to allow signal through

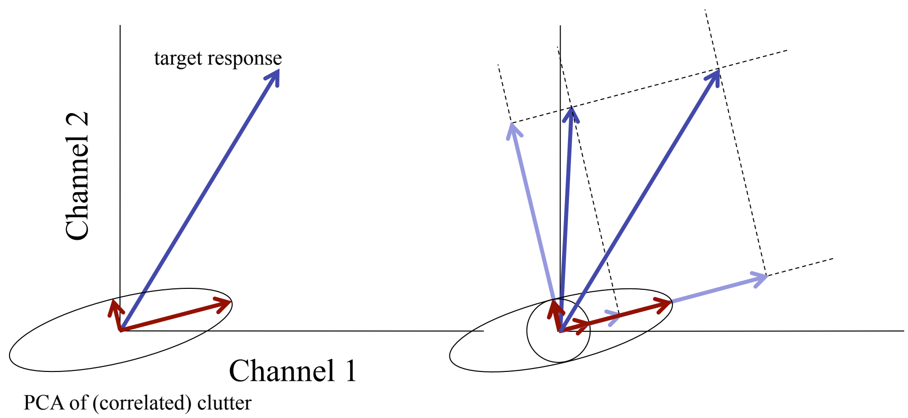
EIGENVECTOR RESEARCH INCORPORATED

Target Projected onto Clutter Directions



EIGENVECTOR RESEARCH INCORPORATED

Scale Target by Clutter

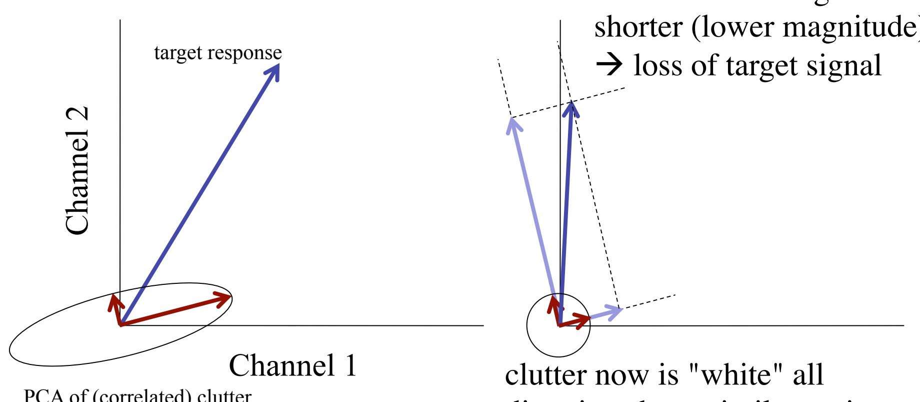


95

EIGENVECTOR RESEARCH INCORPORATED

Whitened Signal

the "whitened" signal is shorter (lower magnitude) → loss of target signal



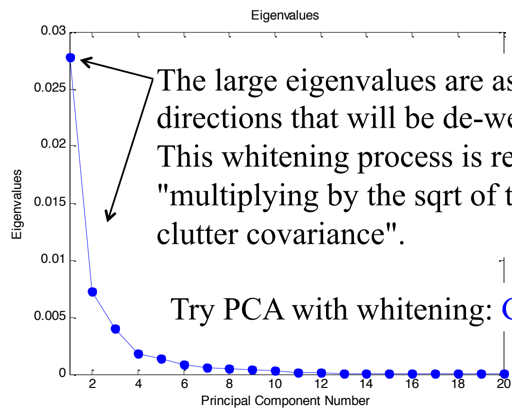
clutter now is "white" all directions have similar variance

96

EIGENVECTOR RESEARCH INCORPORATED

Olive Oil Clutter

Eigenvalue distribution of the within class variance.



The large eigenvalues are associated with directions that will be de-weighted. This whitening process is referred to as "multiplying by the sqrt of the inverse clutter covariance".

Try PCA with whitening: GLS weighting

Based on concepts outlined by Aitken, A., "On Least Squares and Linear Combinations of Observations", *Proceedings of the Royal Society of Edinburgh*, 1935, **55**, 42-48, and used in Maximum Noise Fractions Green AA, Berman M, Switzer P, Craig MD (1988) IEEE Trans Geosci Remote Sens 26:65–74

97

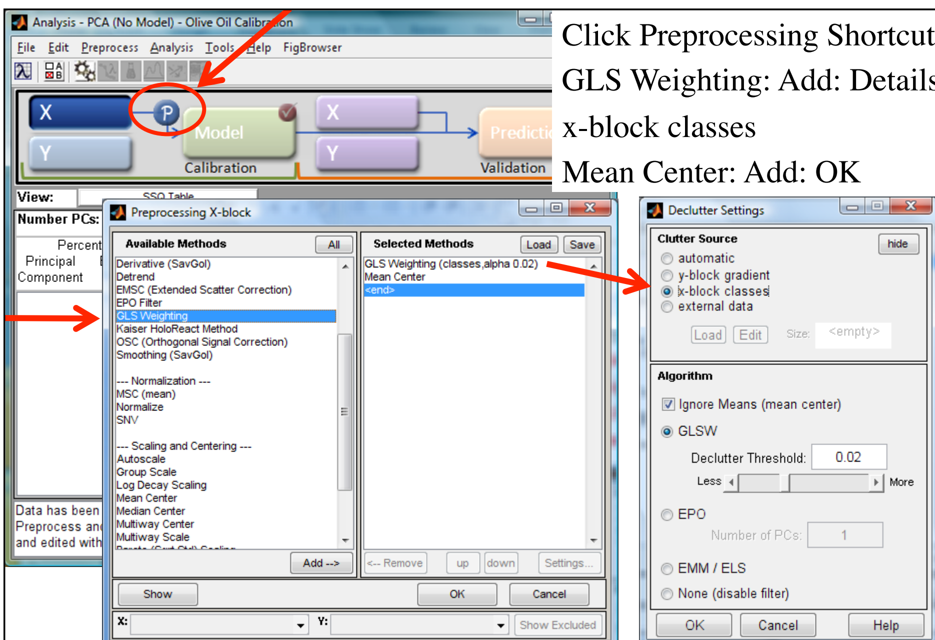
EIGENVECTOR RESEARCH INCORPORATED

Click Preprocessing Shortcut

GLS Weighting: Add: Details:

x-block classes

Mean Center: Add: OK



98

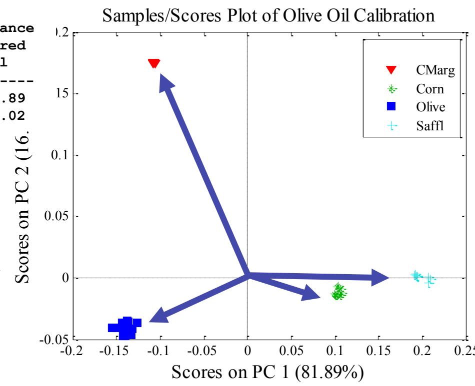
EIGENVECTOR RESEARCH INCORPORATED

PCA of Whitened Spectra

Percent Variance Captured by PCA Model

Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	5.63e-002	81.89	81.89
2	5.34e-003	16.13	98.02

- PCA shows the four classes in the calibration data set are further apart than non-whitened data (very high between class variance) and
- have reduced within class variance



99

Olive Oil Samples

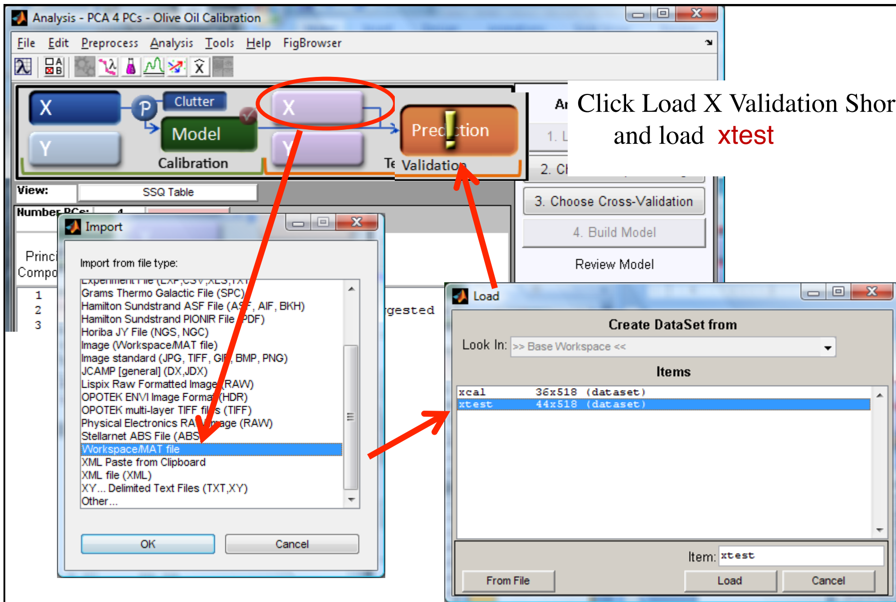
Learning set: xcal Start with this data set

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

Test set: xtest NEW DATA

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 samples	(#37-41)
5, 10, 20, 30 & 40%		
Almond Oil	1 sample	(#42)
Peanut Oil	1 sample	(#43)
Sesame Oil	1 sample	(#44)

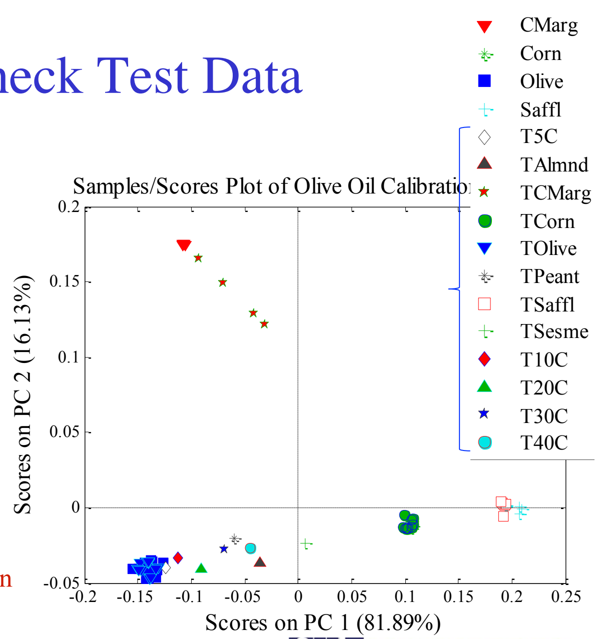
Click Load X Validation Shortcut and load xtest



101

Check Test Data

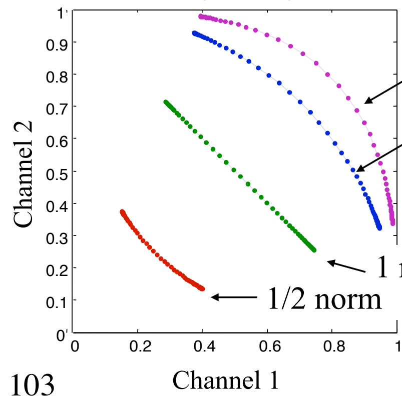
- Test samples for olive oil and corn oil appear to cluster on top of calibration data (good!).
- However, the corn margarine and safflower test samples don't, why?
- Corn margarine suggests a magnitude effect. How to remove the effect of magnitude?
- standard normal variate
- spectral normalization
- multiplicative scatter correction



102

Normalization

- Normalize each row / spectrum
  - 1-norm: normalize to unit AREA (area = 1)
  - 2-norm: normalize to unit LENGTH (vector length = 1)
  - inf-norm: normalize to unit MAXIMUM (max value = 1)



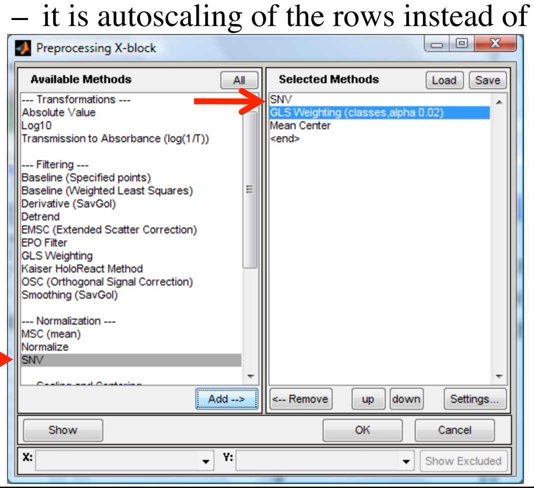
p norm

$$\mathbf{x} = \mathbf{x} \Big/ \left(\sum_{n=1}^{N} |x|_n^p \right)^{1/p}$$

EIGENVECTOR RESEARCH INCORPORATED

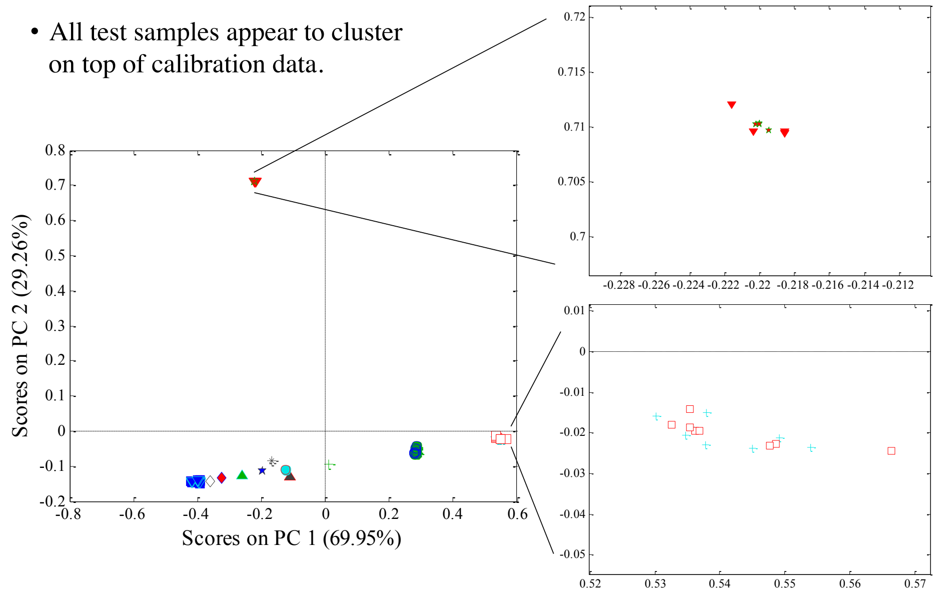
Standard Normal Variate

- Mean-centers each row / spectrum and scales it by it's standard deviation
  - it is autoscaling of the rows instead of the columns



EIGENVECTOR RESEARCH INCORPORATED

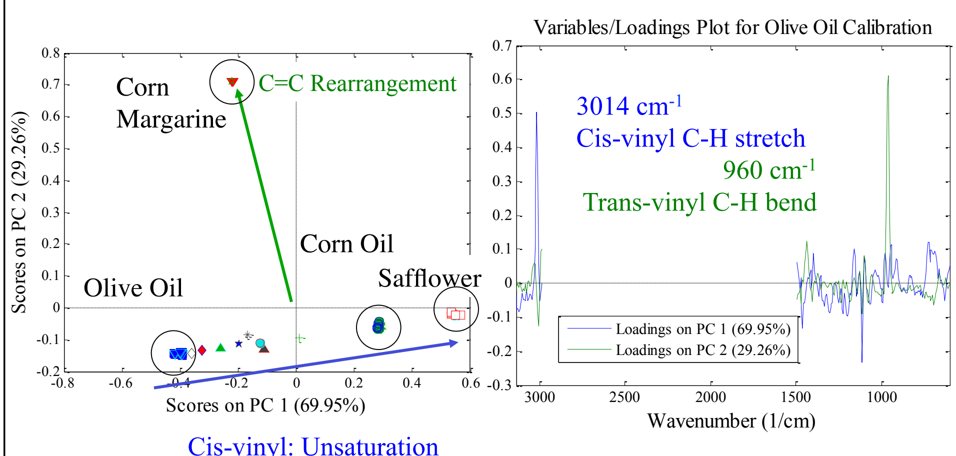
- All test samples appear to cluster on top of calibration data.



105

EIGENVECTOR RESEARCH INCORPORATED

Interpret Scores and Loadings



106

EIGENVECTOR RESEARCH INCORPORATED

GLS Weighting

- GLS Weighting of the spectral data accounted for some of the clutter observed in the spectra, but didn't account for magnitude changes.
- SNV was used to account for magnitude changes.
- The result was
  - clusters were further apart and tighter
  - the ratio of between-class to within-class variance was increased making discrimination easier
    - clusters were so tight and far apart that confidence bounds defining each class could be wider

107

EIGENVECTOR RESEARCH INCORPORATED

Multiplicative Effect in Spectra

- 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 instrument.
- Plotting a measured spectrum versus a reference spectrum (usually the mean) looks linear

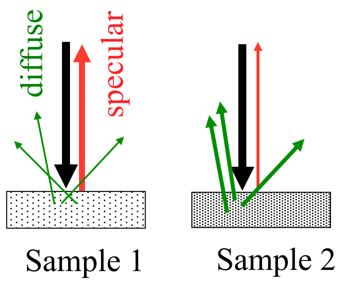
108

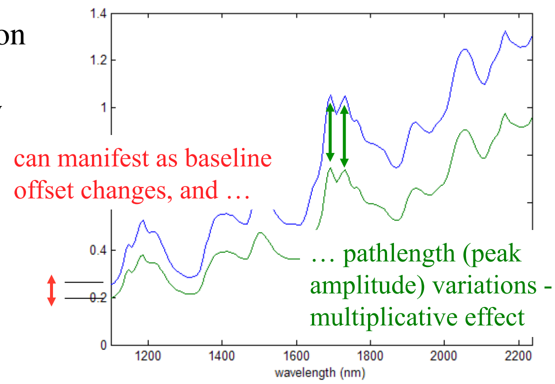
EIGENVECTOR RESEARCH INCORPORATED

Scattering Effects in Reflectance

Caused by variations in:

- Particle mean &distribution
- Sample opacity
- Sampling packing density
- Sample placement

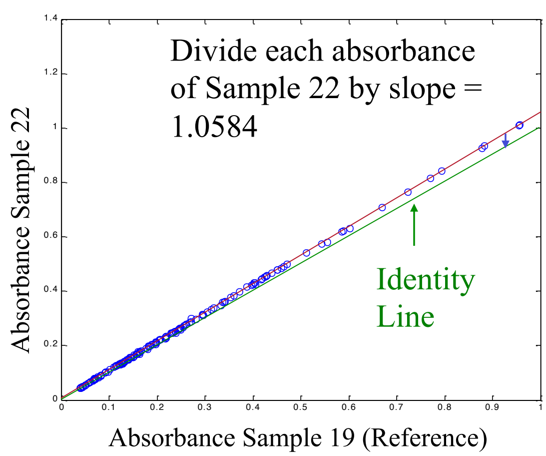




109

EIGENVECTOR RESEARCH INCORPORATED

MSC Multiplicative Signal (Scatter) Correction

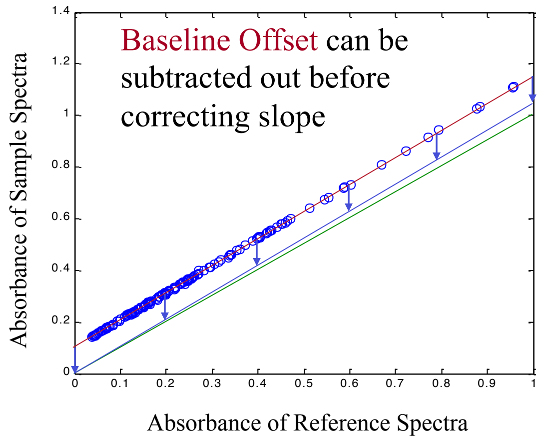


Geladi P, MacDougall D, Martens H., *Appl. Spectrosc.*, **39**(3), 491-500 (1985)

110

EIGENVECTOR RESEARCH INCORPORATED

If there is also an Offset



111

EIGENVECTOR RESEARCH INCORPORATED

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 spectra subtracted when mean centering.

112

EIGENVECTOR RESEARCH INCORPORATED

Extended MSC (EMSC)

- Like MSC, EMSC is used to account for offset and gain (multiplicative effects). Also:
  - clutter by using an extended mixture model
    - using interference spectra or PCA loadings of clutter data
  - instrument artifacts like slope and smile
  - can allow desired target spectra to 'pass the filter'
- The extended mixture model is a classical least squares-like model that is used to explicitly account for clutter using extended least squares.

113

EIGENVECTOR RESEARCH INCORPORATED

EMSC

Provide spectra of:

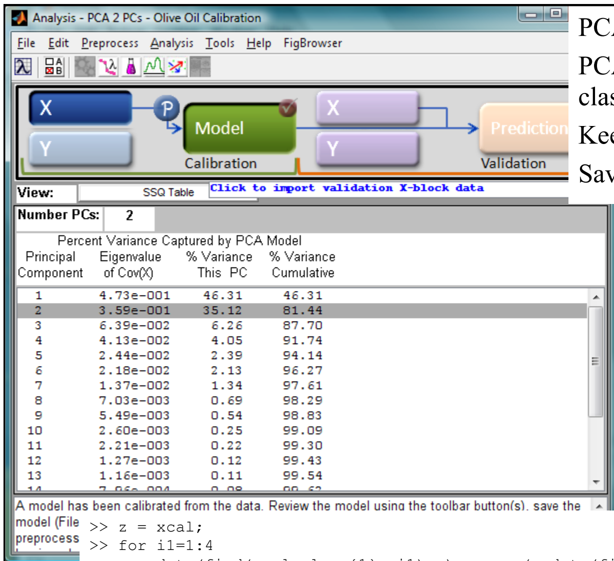
- Known ***target analytes*** **S**
- Polynomial baselines **P**
- Known ***interferences*** **Q**
  - e.g., loadings from a PCA model of clutter
  - the coefficients for each linear effect are estimated using least-squares (indicated by "hat")

$$\mathbf{s}_{2,measured} = \mathbf{s}_{ref}c_{ref} + \mathbf{S}\mathbf{c}_S + \mathbf{P}\mathbf{c}_P + \mathbf{Q}\mathbf{c}_Q \qquad \mathbf{P} = \begin{bmatrix} \mathrm{L} & \boldsymbol{\upsilon}^2 & \boldsymbol{\upsilon} & \mathbf{1} \end{bmatrix}$$

$$\mathbf{s}_{2,corrected} = \left(\mathbf{s}_2 - \mathbf{P}\hat{\mathbf{c}}_P - \mathbf{Q}\hat{\mathbf{c}}_Q\right) / \hat{c}_{ref} \qquad \mathbf{Q} = \text{loadings}$$

114

EIGENVECTOR RESEARCH INCORPORATED



PCA of clutter.

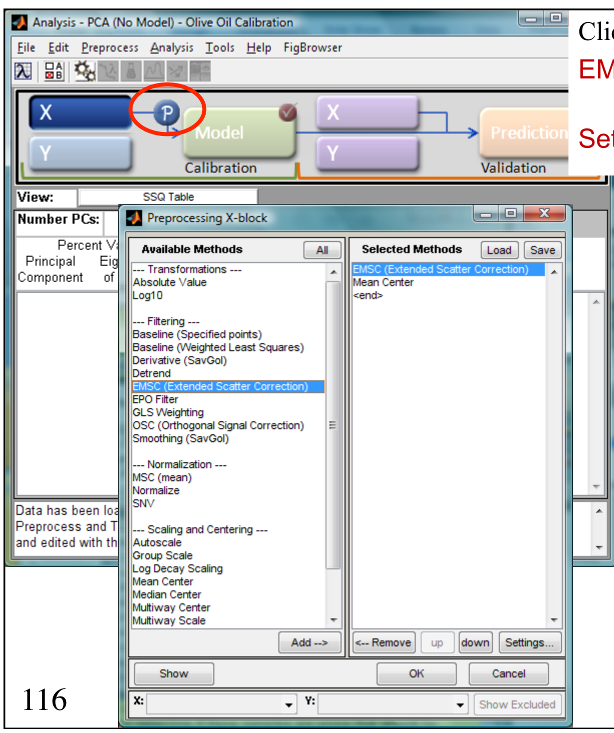
PCA of calibration data with classes centered to class mean.

Keep 2 PCs to model the clutter.

Save model to pcam

```
>> z = xcal;
>> for i1=1:4
      z.data(find(xcal.class{1}==i1),:) = mncn(z.data(find(xcal.class{1}==i1),:));
   end
>> z.description = char(z.description,'Each class center to its own mean.');
>> p = zeros(2,518);
>> p(:,z.include{2}) = pcam.loads{2}';
```

115

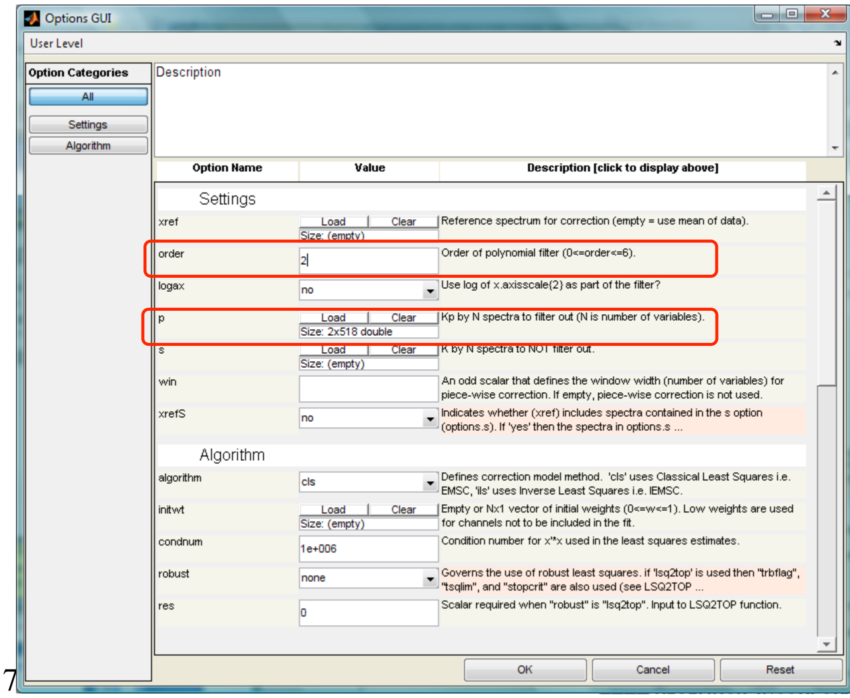


Click Preprocessing Shortcut

EMSC (Extended Multiplicative Scatter Correction)

Settings…

116

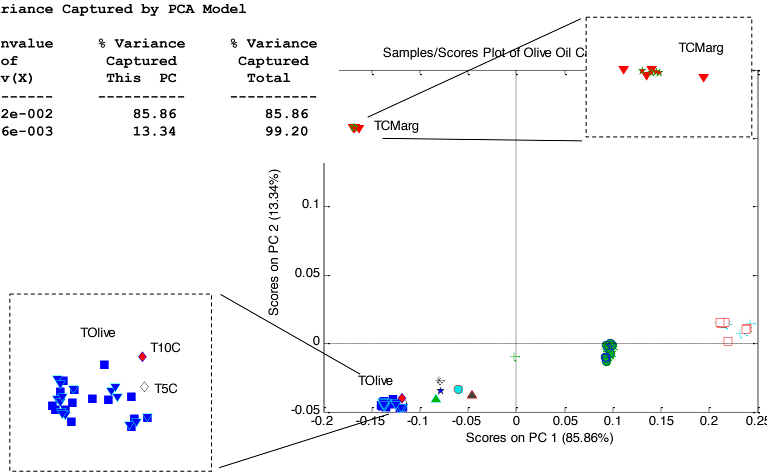


117

EMSC

Percent Variance Captured by PCA Model

Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	2.42e-002	85.86	85.86
2	3.76e-003	13.34	99.20



118

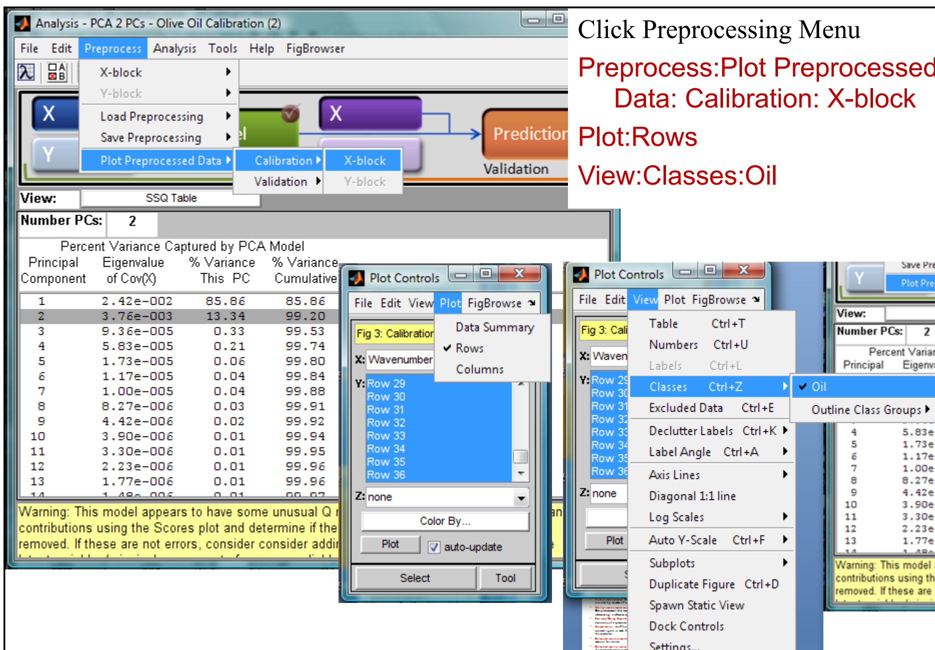
EIGENVECTOR RESEARCH INCORPORATED

EMSC Summary

- EMSC attempts to account for clutter explicitly
  - e.g., model clutter with basis vectors (e.g., PCA loads)
  - analyst takes control of the model
    - requires good use of measurements: clutter and target spectra
    - use what you know!
  - interpretable
    - analyst control is more daunting that simple SavGol and MSC, but
    - results are much more interpretable than 2nd derivative spectra
  - Martens H, Stark E., *J. Pharm. and Biomedical Analysis*, **9**, 625–635 (1991).
  - Helland IS, Naes T, Isaksson T., *Chemom. Intell. Lab. Syst.*, **29**, 233–241 (1995).
  - Martens H, Nielsen JP, Engelsen SB., *Anal. Chem.*, **75**(3), 394–404 (2003).
  - Gallagher NB, Blake TA, Gassman PL, *J. Chemometr.*, **19**(5-7), 271-281 (2005).

119

EIGENVECTOR RESEARCH INCORPORATED



Click Preprocessing Menu

Preprocess:Plot Preprocessed Data: Calibration: X-block

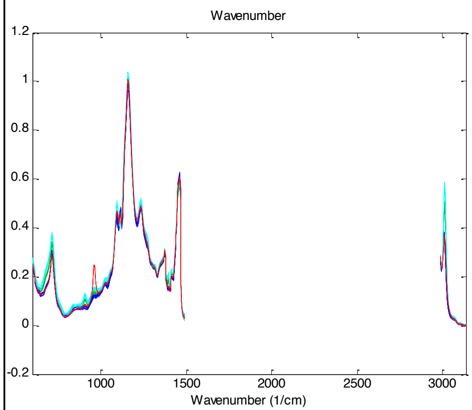
Plot:Rows

View:Classes:Oil

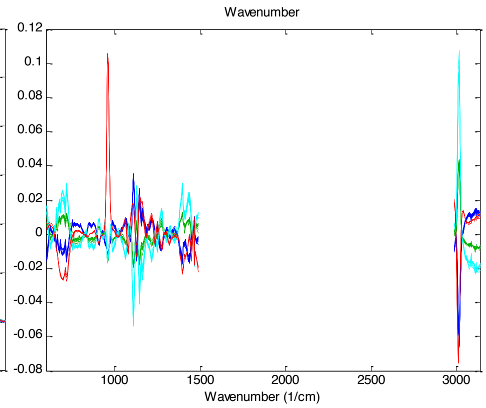
120

EIGENVECTOR RESEARCH INCORPORATED

Original data



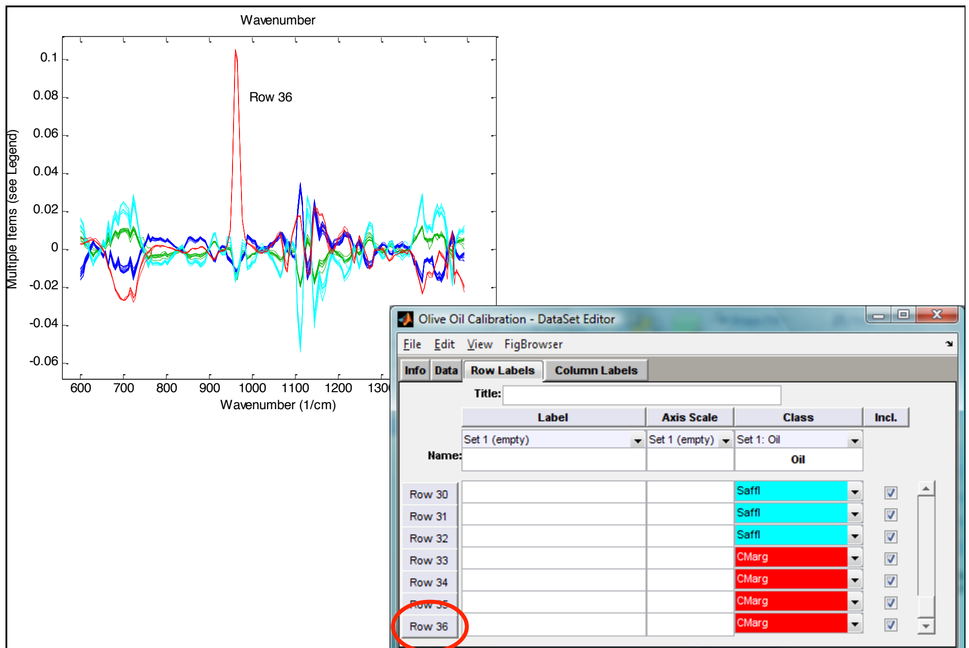
EMSC processed data, mean-centered



(colored by class)

121

EIGENVECTOR RESEARCH INCORPORATED



122

EIGENVECTOR RESEARCH INCORPORATED

Outline

- Introduction
- PCA Review
- PLS Regression Review
- Advanced Preprocessing
- Variable Selection
  - why do it?
  - use what you know!
  - iPLS
- Summary

123

EIGENVECTOR RESEARCH INCORPORATED

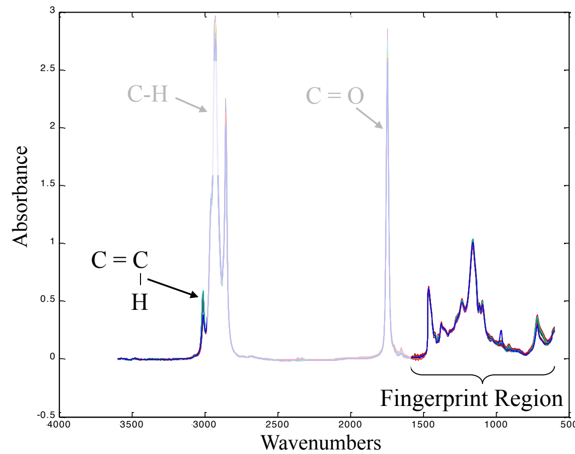
Why Variable Selection?

- Improvement of the model
  - Remove irrelevant, unreliable or noisy variables (clutter)
  - Improve predictions
  - Improve statistical properties
- Interpretation
  - Obtain a model that is easier to understand
- Costs
  - Use fewer measurements to replace expensive or time-consuming one
- Development of fast instruments/routines for on-line control
  - Find wavelength ranges for a filter-based instrument

124

EIGENVECTOR RESEARCH INCORPORATED

Already performed variable selection based on *a posteriori* knowledge…



125

Variable Selection Methods

- ***a priori***
  - Choose measurements
- ***a posteriori***
  - Use chemical/physical insight
- **Model based**
  - Look at loadings
- **"Random based"**
  - Genetic algorithms
  - Simulated annealing
- **"Spectral"**
  - i-PLS
  - fullsearch
- **Classical**
  - Forward, backward selection
  - Best subset selection
  - Significance tests
  - Significance based on Jack-knife
  - GOLPE
- **Other**
  - Pure variables
  - Principal variables
  - Iterative weighting with regression vector
  - …

(see the Variable Selection Course at EigenU)

126

Variable Selection Methods

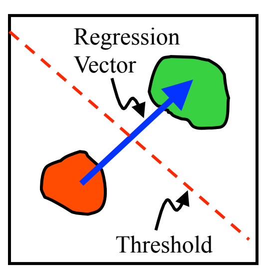
- How to choose which method?!?
- Different methods work in different situations
- Interval-PLS is a good "example" method to understand the considerations of variable selection. Simple to implement and use.
- Can be used on the Olive Oil data set, but first need to define PLS-DA

127 EIGENVECTOR RESEARCH INCORPORATED

Partial Least Squares-Discriminate Analysis (PLS-DA)

- Use logicals (0,1) in Y-block to indicate if sample belongs to a class or not.
- Develop PLS model to predict class block
- Thresholds must be set between 0 and 1 to indicate if new samples are a member of each class...

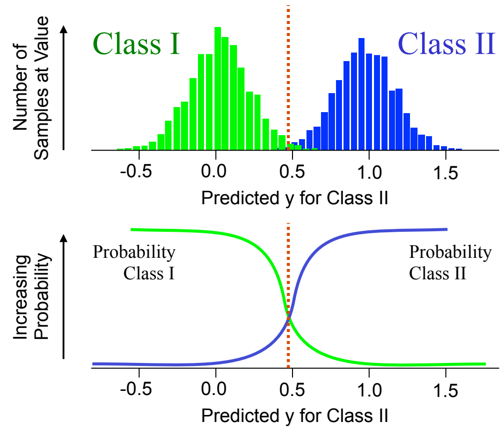
  Can use Bayes theorem to set threshold and include prior probability of each class



128 EIGENVECTOR RESEARCH INCORPORATED

Thresholds in PLS-DA

Observed distribution of predictions can be handled in a straight-forward Bayesian way



129

EIGENVECTOR RESEARCH INCORPORATED

PLS-DA for Olive Oil Data

- PLS-DA tends to capture variance which is useful in separating classes and ignoring variance within a class.
  - goal: maximize inter-class variance while minimizing intra-class variance
- For Olive Oils it seems reasonable to discriminate Corn Margarine from all the others first.
  - Other classes can be separated in turn
  - Two classes: Corn margarine and Everything else
    - this was evident based on the previous exploratory analysis

130

EIGENVECTOR RESEARCH INCORPORATED

Re-Include all variables

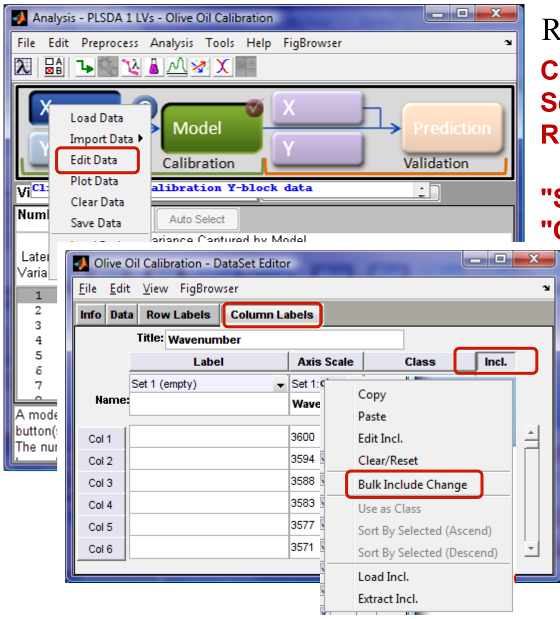
Click "X" >> Edit Data

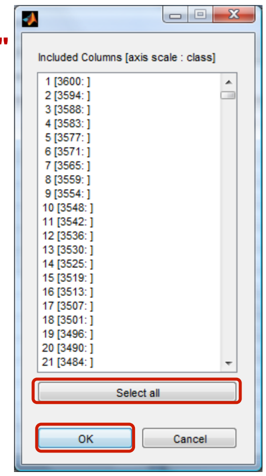
Select Column Labels tab

Right Click Incl. >> Bulk Include Change

"Select all"

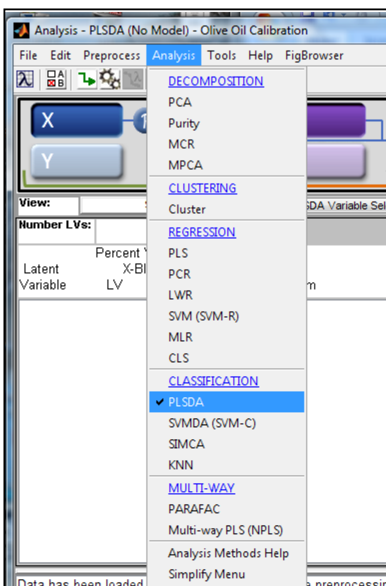
"OK"

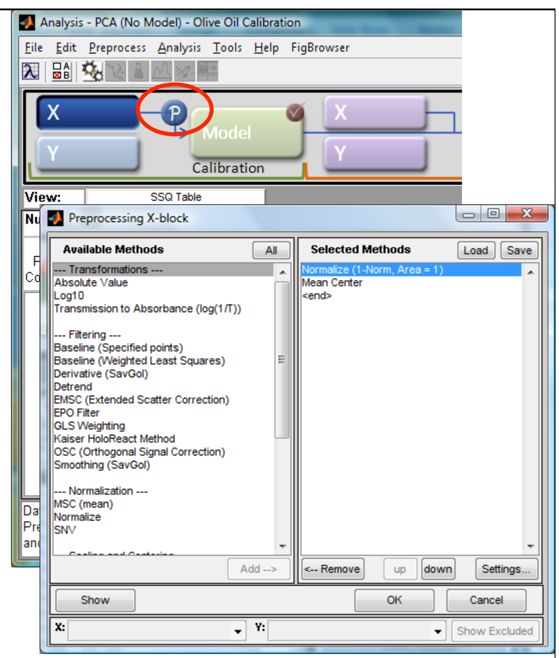




EIGENVECTOR RESEARCH INCORPORATED

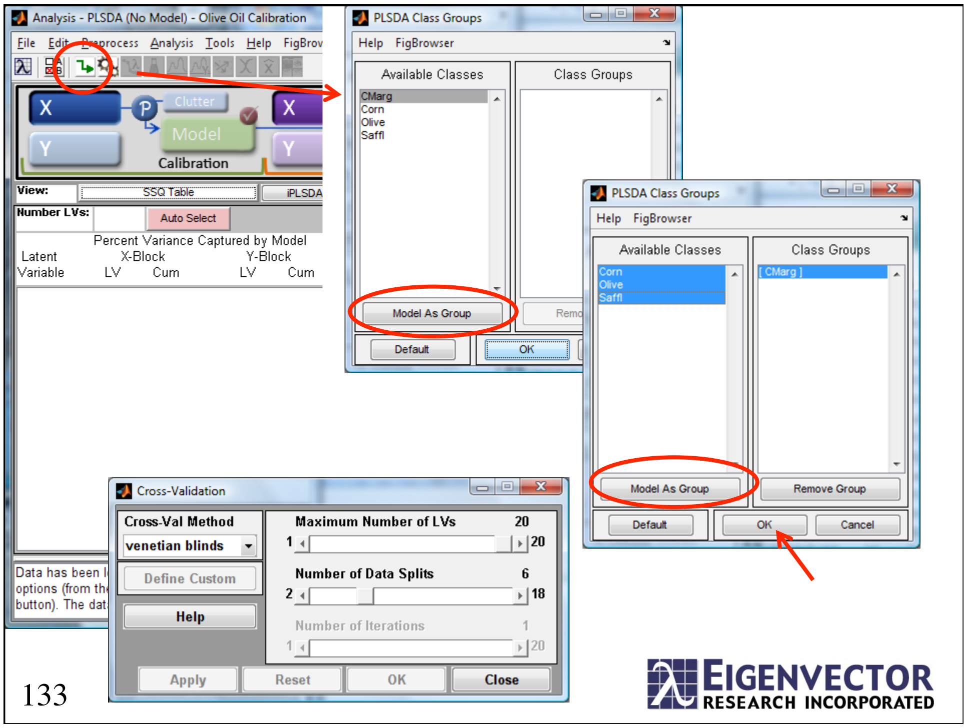
131

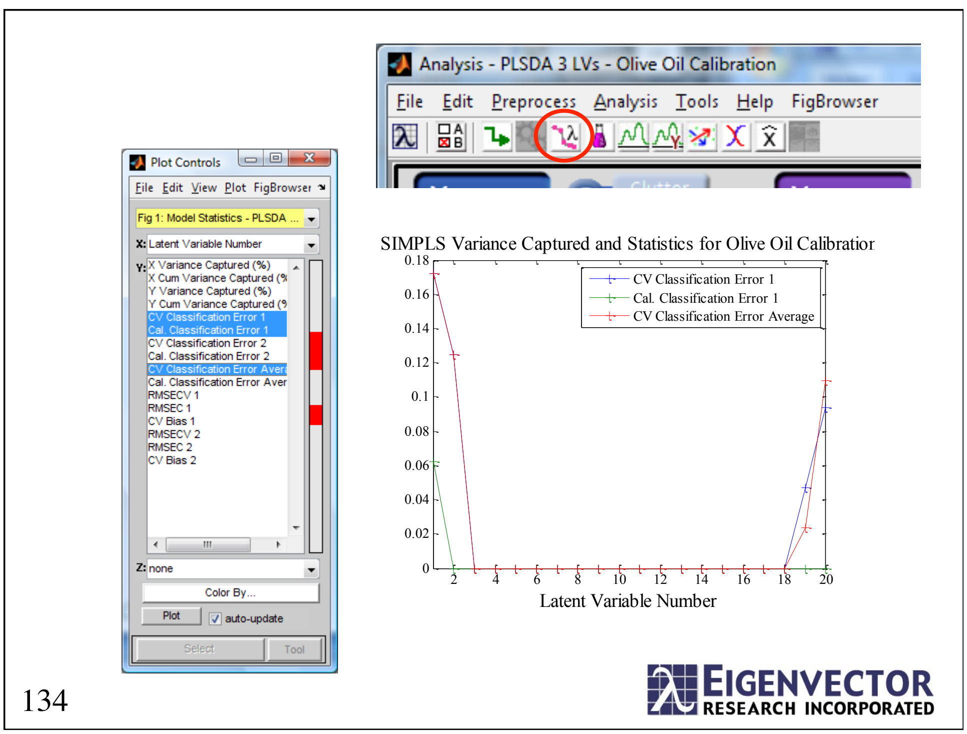


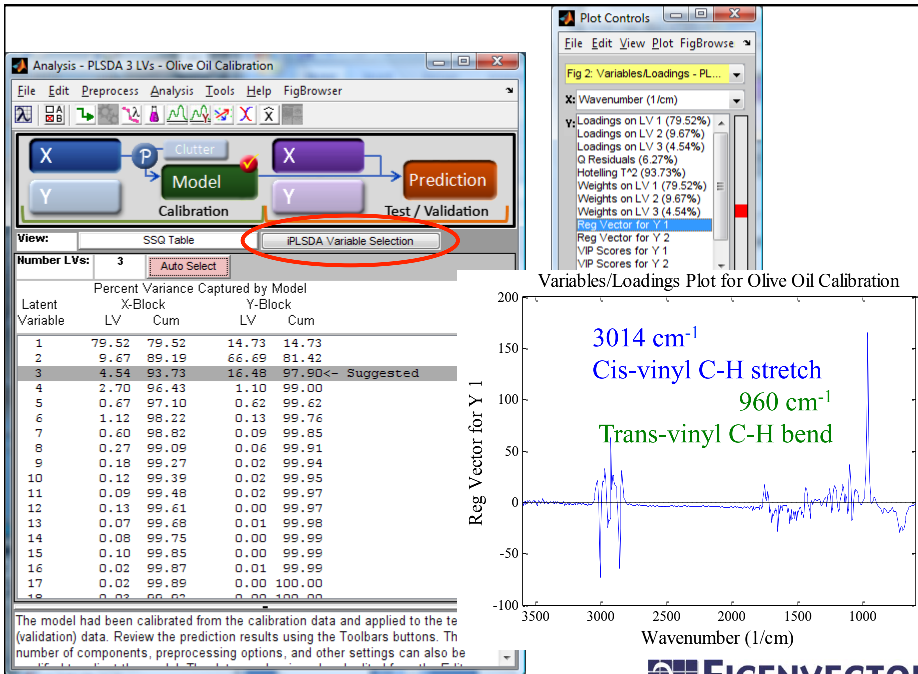


EIGENVECTOR RESEARCH INCORPORATED

132

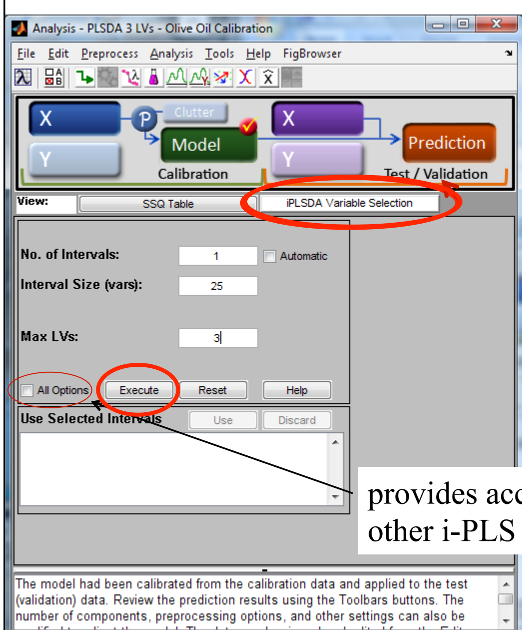

Analysis - PLSDA (No Model) - Olive Oil Calibration
File Edit Preprocess Analysis Tools Help FigBrow
Clutter
Model
Calibration
View:
SSQ Table
iPLSDA
Number LVs:
Auto Select
Percent Variance Captured by Model
Latent Variable
X-Block LV Cum
Y-Block LV Cum
Data has been l
options (from th
button). The dat
PLSDA Class Groups
Help FigBrowser
Available Classes
Class Groups
CMarg
Corn
Olive
Saffl
Model As Group
Default
OK
PLSDA Class Groups
Help FigBrowser
Available Classes
Class Groups
Corn
Olive
Saffl
[CMarg]
Model As Group
Remove Group
Default
OK
Cancel
Cross-Validation
Cross-Val Method
venetian blinds
Define Custom
Help
Maximum Number of LVs 20
1 20
Number of Data Splits 6
2 18
Number of Iterations 1
1 20
Apply
Reset
OK
Close
133
EIGENVECTOR
RESEARCH INCORPORATED


Analysis - PLSDA 3 LVs - Olive Oil Calibration
File Edit Preprocess Analysis Tools Help FigBrowser
Plot Controls
File Edit View Plot FigBrowser
Fig 1: Model Statistics - PLSDA ...
X: Latent Variable Number
Y: X Variance Captured (%)
X Cum Variance Captured (%
Y Variance Captured (%)
Y Cum Variance Captured (%
CV Classification Error 1
Cal. Classification Error 1
CV Classification Error 2
Cal. Classification Error 2
CV Classification Error Avera
Cal. Classification Error Aver
RMSECV 1
RMSEC 1
CV Bias 1
RMSECV 2
RMSEC 2
CV Bias 2
Z: none
Color By...
Plot
auto-update
Select
Tool
SIMPLS Variance Captured and Statistics for Olive Oil Calibration
CV Classification Error 1
Cal. Classification Error 1
CV Classification Error Average
0.18
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
2
4
6
8
10
12
14
16
18
20
Latent Variable Number
134
EIGENVECTOR
RESEARCH INCORPORATED



135

EIGENVECTOR RESEARCH INCORPORATED



Click "IPLSDA Variable Selection" button
Enter 25 for Interval Size
Enter 3 for max LVs
Click "Execute"

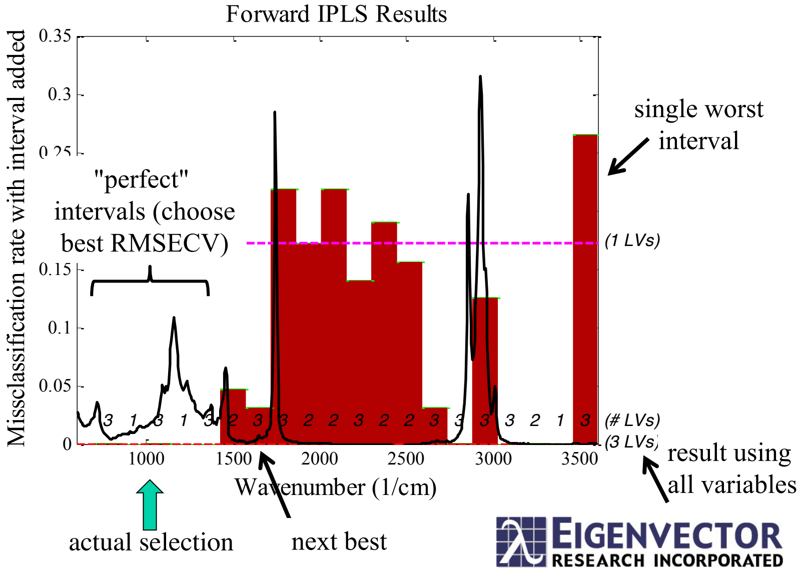
Interval Size defines how many variables to group together.
Spectral data: use size>1
Non-contiguous data: use size=1 (single variable intervals)

provides access to other i-PLS settings

136

EIGENVECTOR RESEARCH INCORPORATED

First Interval Result

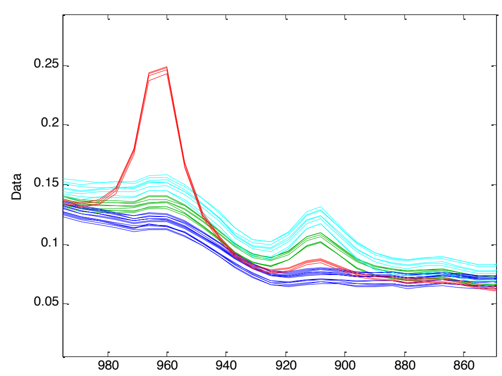


137

EIGENVECTOR RESEARCH INCORPORATED

Why the Low-Frequency Intervals? Information Content

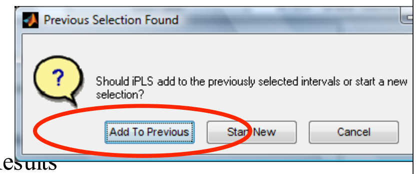
- Best intervals contain useful signal from all four classes.
- Original data shown; Preprocessing only makes this separation better!

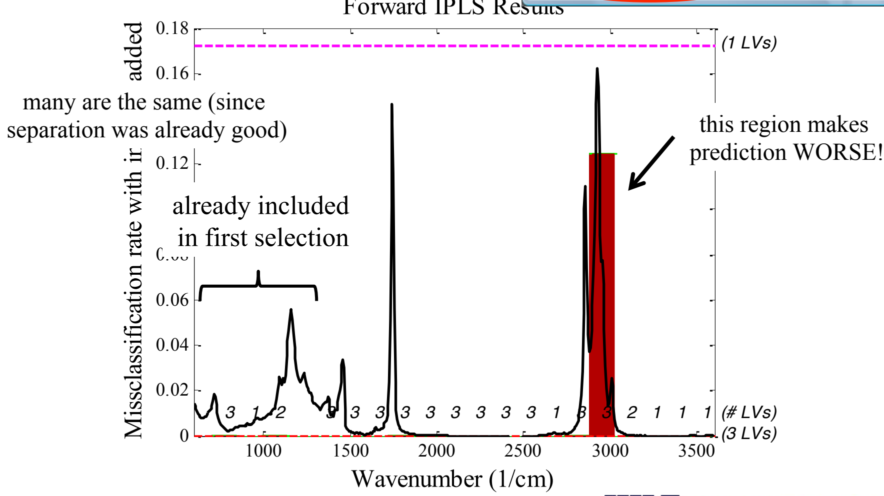


138

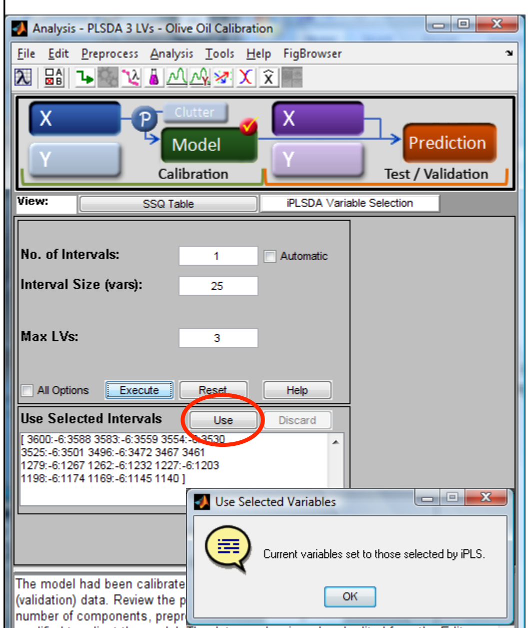
EIGENVECTOR RESEARCH INCORPORATED

Repeat "Execute" and "Add To Previous"



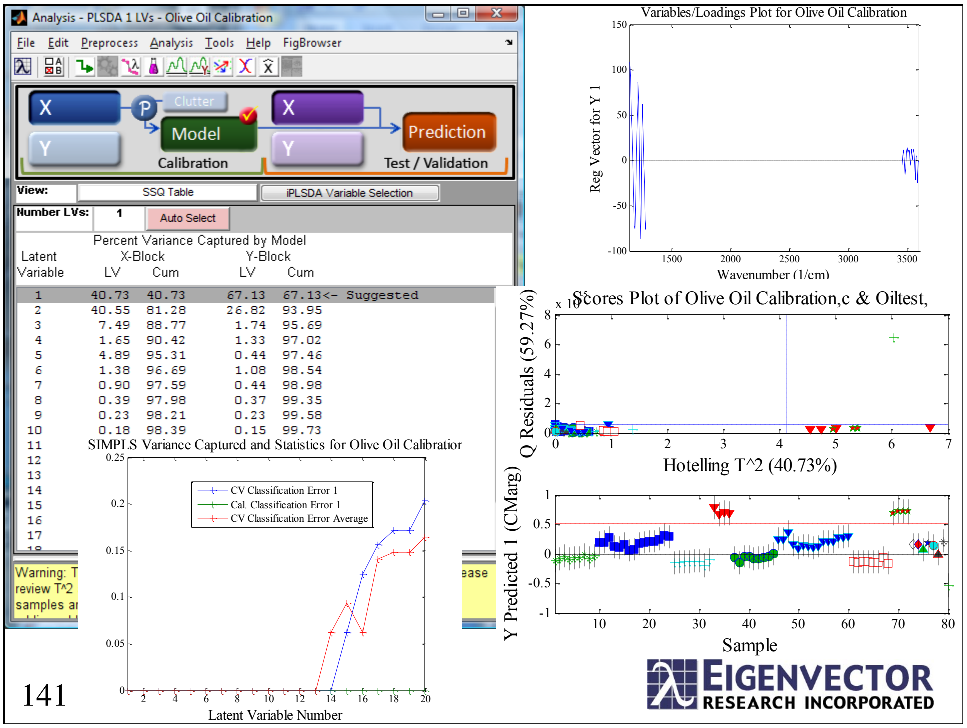


139



Click "Use" button
Click "OK"
Click "SSQ Table" button
Click "Model"

140



Number of Intervals

- Can choose a pre-set number of intervals to find
- Can also use "Automatic" to continue selecting intervals until RMSECV/misclassification does not improve
- This is **not** the same as exhaustive combinatorial search (fullsearch). It is sequential (choose one, "lock" it in, choose a second, "lock" it in…)
- For very complex data, may not give actual "best" windows, but probably not a bad one.

142

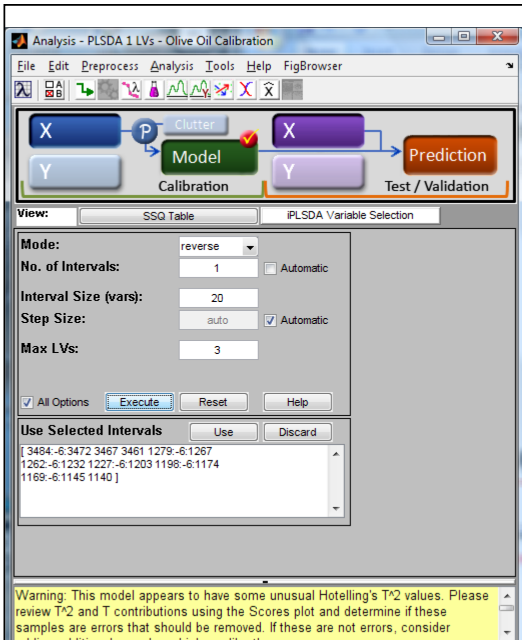
EIGENVECTOR RESEARCH INCORPORATED

What is the Result?

- For PLSDA, lower RMSECV should indicate better class separation in predicted Y values
- Selecting additional intervals gives little improvement in RMSECV (on this data)
- Use ONLY first selected interval and build new model…

143

EIGENVECTOR RESEARCH INCORPORATED



144

EIGENVECTOR RESEARCH INCORPORATED

Outline

- Introduction
- PCA Review
- PLS Regression Review
- Advanced Preprocessing
- Variable Selection
- Summary

145

EIGENVECTOR RESEARCH INCORPORATED

Summary

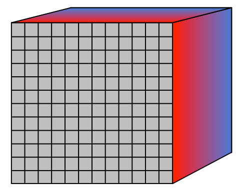
- Data analysis requires knowledge of
  - the system, physics, chemistry and math → ~~black box~~
- Advanced Preprocessing
  - uses knowledge of the clutter (GLS, ELS, etc.)
- Variable Selection
  - choose variables that are most predictive

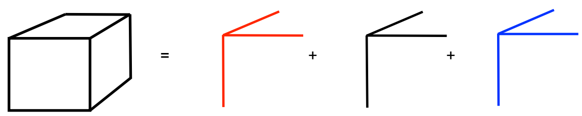
146

EIGENVECTOR RESEARCH INCORPORATED

If time …

- Introduce
  - multivariate image analysis (hyperspectral image analysis)
  - multiway analysis





147

EIGENVECTOR RESEARCH INCORPORATED

Outline

- Introduction
- PCA Review
- PLS Regression Review
- Advanced Preprocessing
- Variable Selection
- Introduction to Multivariate Image Analysis and Multi-way Analysis (if time)
- Summary

148

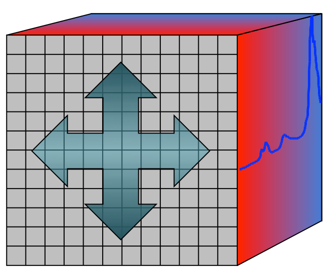
EIGENVECTOR RESEARCH INCORPORATED

Multivariate Images

A data array of *dimension three* (or more) where the first two dimensions are *spatial* and the last dimension(s) is a function of another variable.

149

Multivariate Images



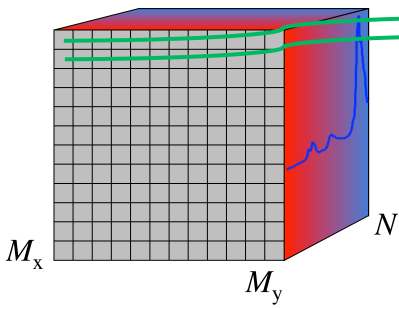
Spectral Information between channels (chemical information)

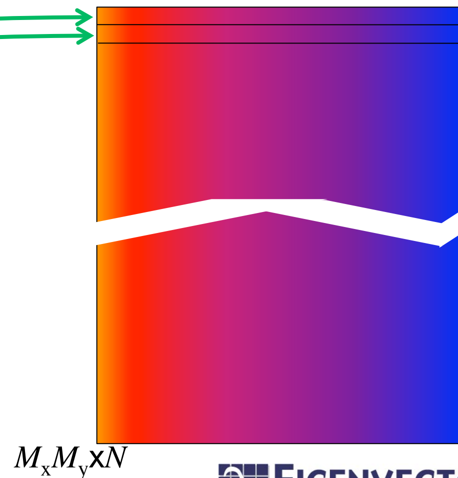
Spatial Information between pixels

Spatial distribution of chemical analytes, physical features, and other properties

150

Reshaping Images for Analysis

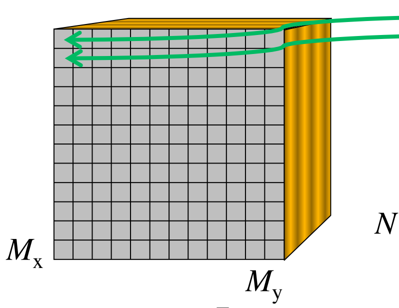




Reshape $\underline{\mathbf{X}}$ (M_xxM_yxN) to $\mathbf{X}$ ($M_x M_y$xN) matrix and subject to two-way analysis e.g, PCA

EIGENVECTOR RESEARCH INCORPORATED

PCA for Multivariate Image Analysis



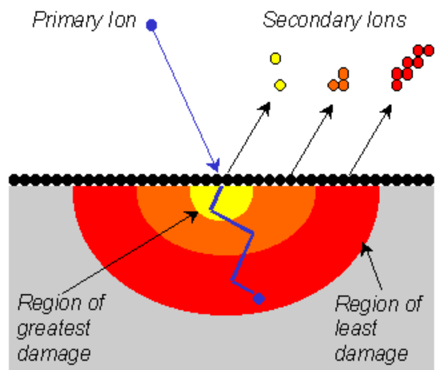
Model $\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$
Reshape $\mathbf{T}$ ($M_x M_y$xK) to $\underline{\mathbf{T}}$ (M_xxM_yxK) and examine the scores images

PCA is an example of MIA. *Techniques and Applications of Hyperspectral Image Analysis*, Grahn, H. F.; Geladi, P., Eds. John Wiley & Sons: West Sussex, England (2007)

EIGENVECTOR RESEARCH INCORPORATED

Example: TOF-SIMS

- Time-of-Flight Secondary-Ion-Mass Spectrometry
  - common surface analysis technique
  - mass spectrum generated for each pixel



Thanks to Physical Electronics!

153

EIGENVECTOR RESEARCH INCORPORATED

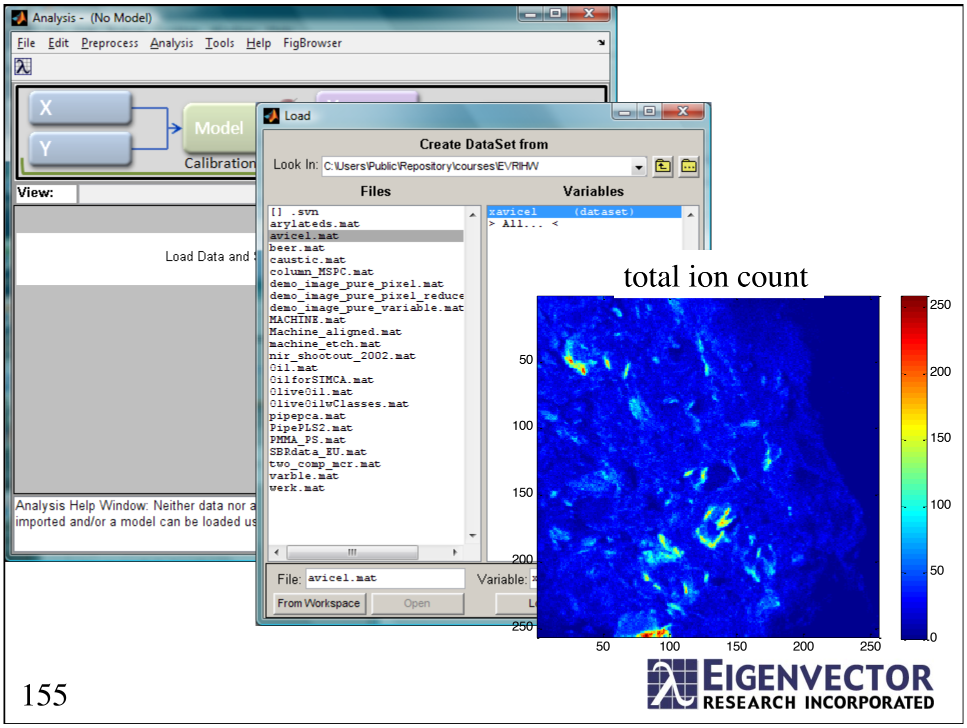
TOF-SIMS of Time Release Drug Delivery System

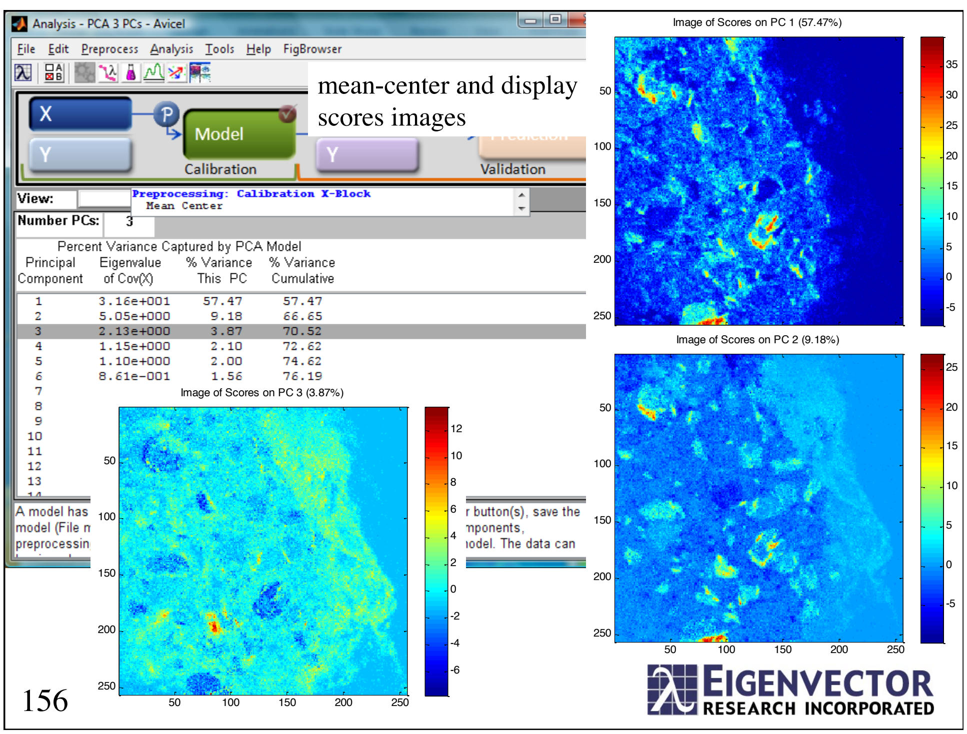
- Multi-layer drug beads serve as controlled release system
- TOF-SIMS of cross section of bead
- Evaluate the integrity of the layers and distribution of ingredients

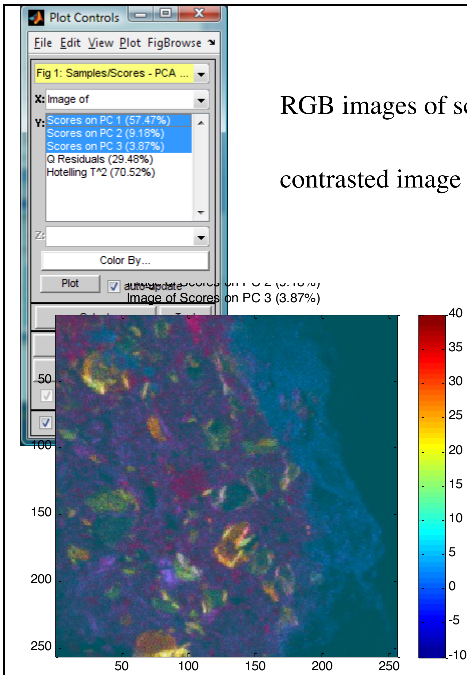
A.M. Belu, M.C. Davies, J.M. Newton and N. Patel, "TOF-SIMS Characterization and Imaging of Controlled-Release Drug Delivery Systems," *Anal. Chem.*, **72**(22), 5625–5638 (2000).
Gallagher, N.B., Shaver, J.M., Martin, E.B., Morris, J., Wise, B.M. and Windig, W., "Curve resolution for images with applications to TOF-SIMS and Raman", *Chemometr. Intell. Lab.*, **73**(1), 105–117 (2003).

154

EIGENVECTOR RESEARCH INCORPORATED

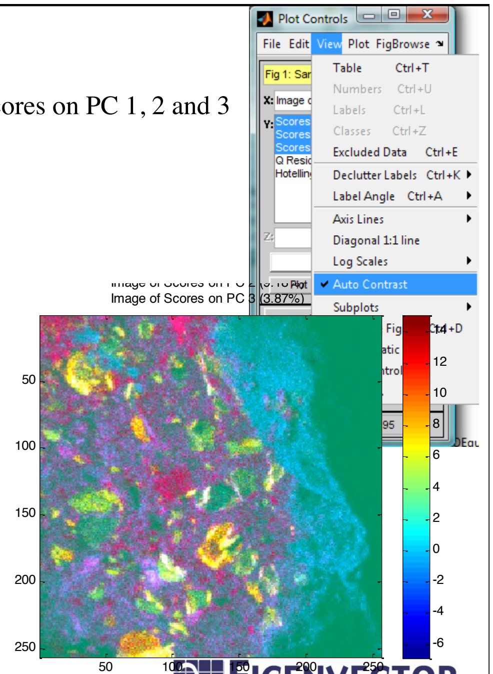

Analysis - (No Model)
File Edit Preprocess Analysis Tools Help FigBrowser
Load
Create DataSet from
Look In: C:\Users\Public\Repository\courses\EVRIHW
Files
Variables
avicel (dataset)
File: avicel.mat
From Workspace
Open
total ion count
155
EIGENVECTOR RESEARCH INCORPORATED


Analysis - PCA 3 PCs - Avicel
File Edit Preprocess Analysis Tools Help FigBrowser
mean-center and display scores images
Calibration
Validation
Preprocessing: Calibration X-Block
Mean Center
Number PCs: 3
Percent Variance Captured by PCA Model
Principal Component | Eigenvalue of Cov(X) | % Variance This PC | % Variance Cumulative
1 3.16e+001 57.47 57.47
2 5.05e+000 9.18 66.65
3 2.13e+000 3.87 70.52
4 1.15e+000 2.10 72.62
5 1.10e+000 2.00 74.62
6 8.61e-001 1.56 76.19
Image of Scores on PC 1 (57.47%)
Image of Scores on PC 2 (9.18%)
Image of Scores on PC 3 (3.87%)
156
EIGENVECTOR RESEARCH INCORPORATED



RGB images of scores on PC 1, 2 and 3

contrasted image



157

Minimum Noise Factors (MNF)

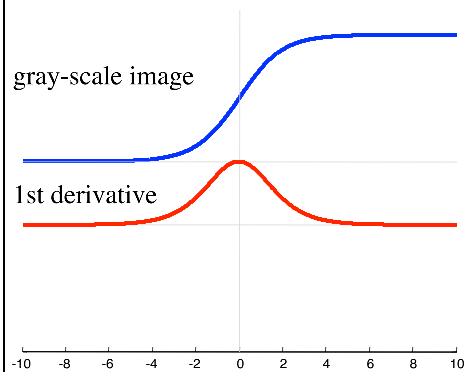
- MNF attempts find directions in the data that maximize the signal-to-clutter.

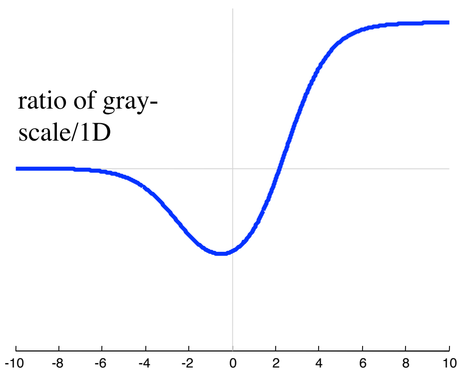
$$\max_{\mathbf{v}_i \neq 0} \left(\frac{\mathbf{v}_i^T \mathbf{\Sigma}_X \mathbf{v}_i}{\mathbf{v}_i^T \mathbf{\Sigma}_C \mathbf{v}_i} \right) \qquad \text{the objective function}$$

- Result is a PCA-like eigenvector problem
- In maximum autocorrelation factors (MAF) clutter is the first difference image (difference between near-by pixels)

158

MAF





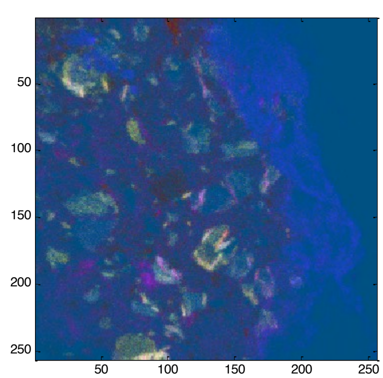
MAF finds locations in the image where the ratio of gray-scale to first derivative is a maximum

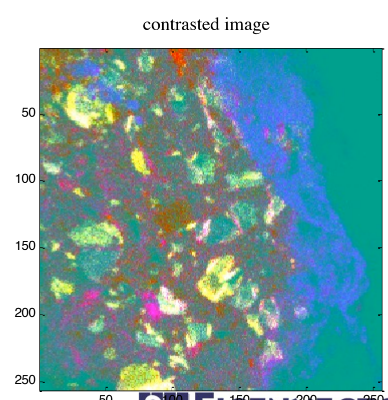
159

EIGENVECTOR RESEARCH INCORPORATED

MAF Results

RGB images of MAF scores on 1, 2 and 3



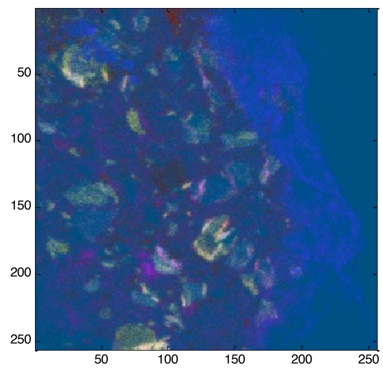


160

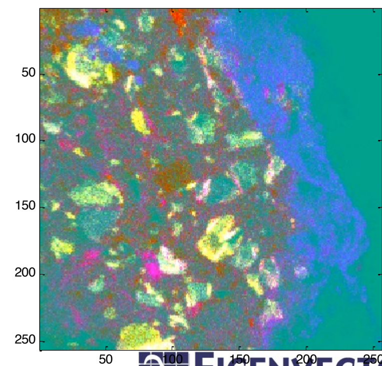
EIGENVECTOR RESEARCH INCORPORATED

PCA w/ GLS Weighting for ~MAF

RGB images of PCA w/ GLS weighting scores on 1, 2 and 3. Similar to MAF results. Objective function ~similar, but PCA scores and loadings orthogonal.



contrasted image



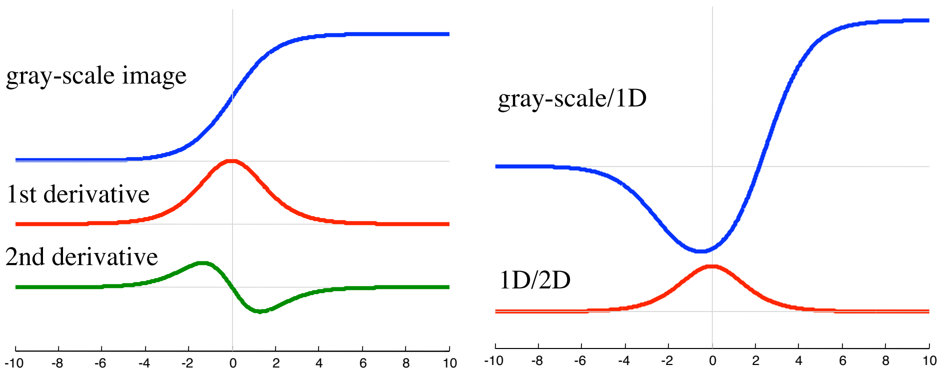
161

Maximum Difference Factors (MDF)

- In MDF the signal covariance corresponds to the first derivative across the spatial dimensions.
  - in MAF the first difference is the clutter
- The clutter corresponds to the second derivative across the spatial dimensions.
- Gives a multivariate analysis estimate of edges in an image.
  - analogous method available for GLS weighting w/ PCA

162

MDF

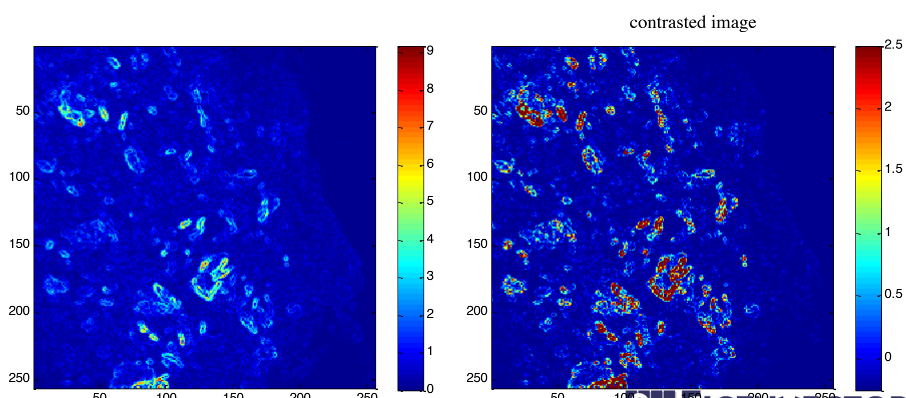


MDF finds locations in the image where the ratio of first to second derivative is a maximum

163

MDF Results

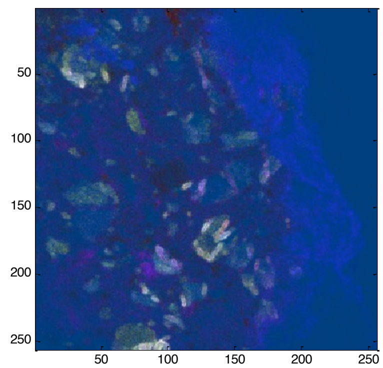
Scaled image of MDF scores on 1
sqrt(dx^2 + dy^2)



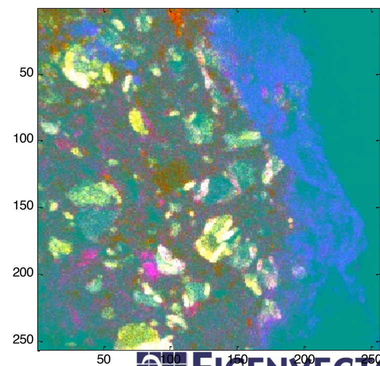
164

MAF+MDF Results

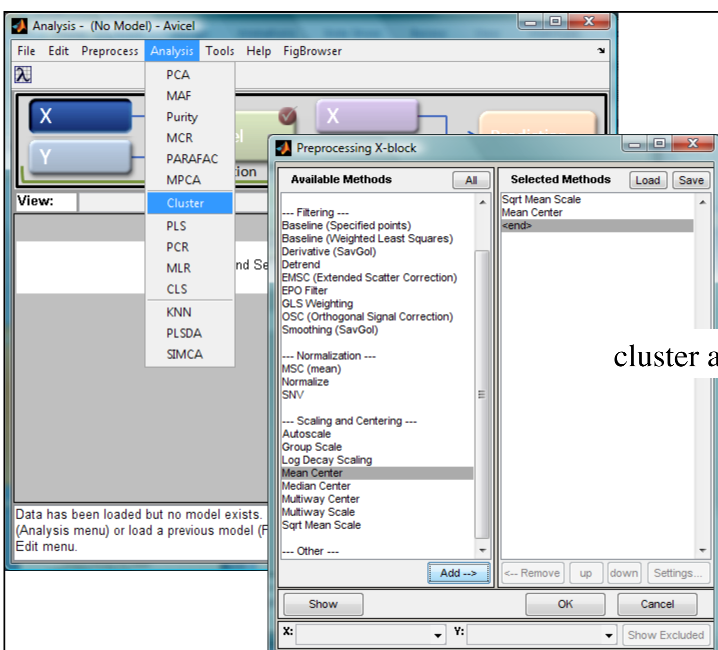
RGB image of MAF scores 1, 2 and 3 + MDF scores on 1
sqrt(dx^2 + dy^2)



contrasted image

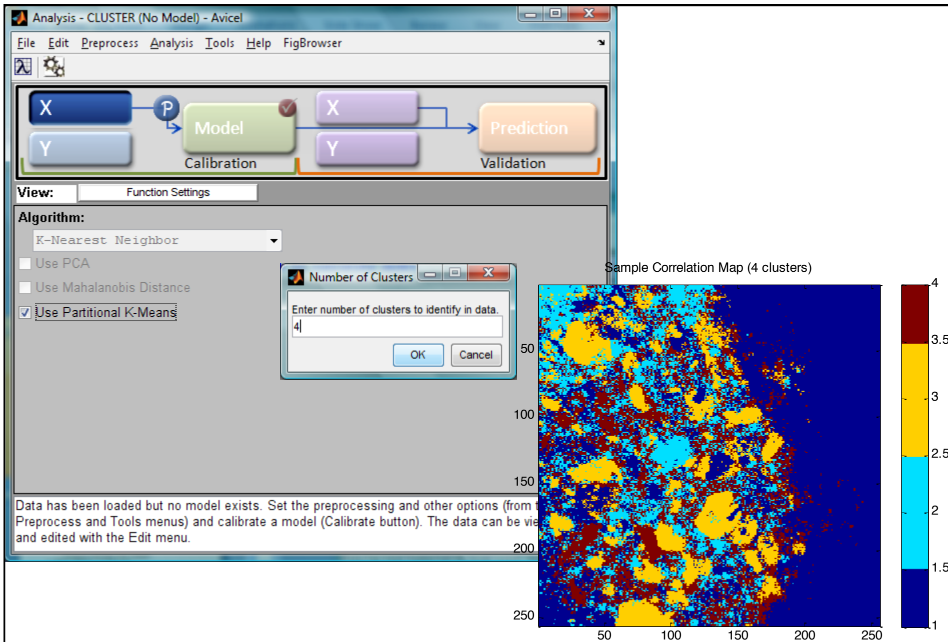


EIGENVECTOR RESEARCH INCORPORATED



cluster analysis

EIGENVECTOR RESEARCH INCORPORATED



167

EIGENVECTOR RESEARCH INCORPORATED

MIA …

- Much more to MIA
  - linked scores plots and density plots
    - interactive exploration of the image(s)
  - image SIMCA and PLS-DA
    - classification
  - curve resolution
    - chemical identification and mapping
  - image statistical process control (ISPC) for multivariate statistical process control (MSPC)
  - …

168

EIGENVECTOR RESEARCH INCORPORATED

Outline

- Introduction
- Advanced Preprocessing
  - Clutter and characterizing clutter
  - Generalized least squares weighting
  - Extended multiplicative scatter correction
  - Interval PLS (iPLS)
  - Model Robustness
- Multivariate image analysis
- Multi-way Analysis
- Summary

EIGENVECTOR RESEARCH INCORPORATED

Why is Clutter Bad?

- What is clutter and how does clutter effect the measured signal?
- Use FT-IR spectra and pattern recognition to distinguish authentic olive oil from counterfeit or adulterated olive oil.

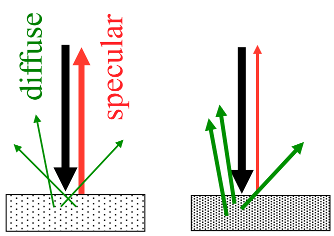
170

EIGENVECTOR RESEARCH INCORPORATED

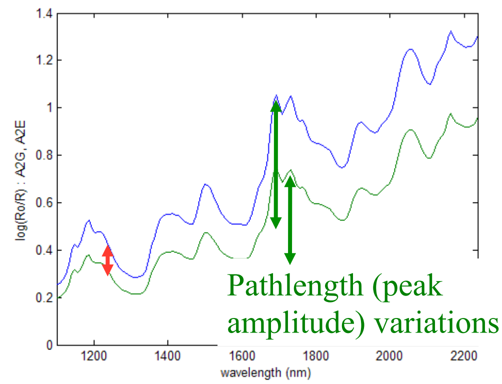
Sources of Clutter: Scattering Effects in Reflectance

Caused by variations in:

- Particle size (mean & distribution)
- Sample opacity
- Sampling packing density
- Sample placement



Sample 1 Sample 2



Baseline offset changes

171

EIGENVECTOR RESEARCH INCORPORATED

Olive Oil Samples

Learning / Calibration Set:

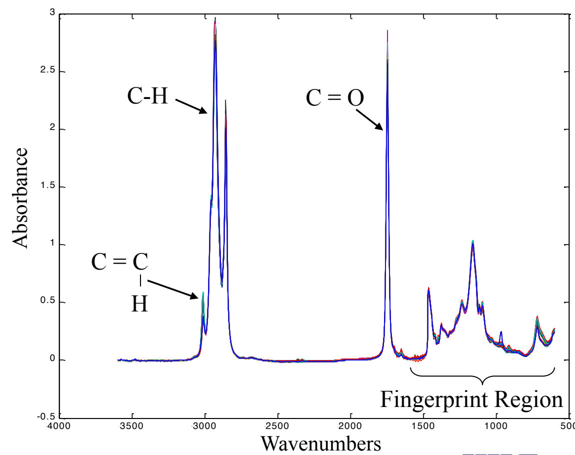
Corn Oil	9 samples	(#1-9)
Olive Oil	15 samples	(#10-24)
Safflower Oil	8 samples	(#25-32)
Corn Margarine	4 samples	(#33-36)

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

172

EIGENVECTOR RESEARCH INCORPORATED

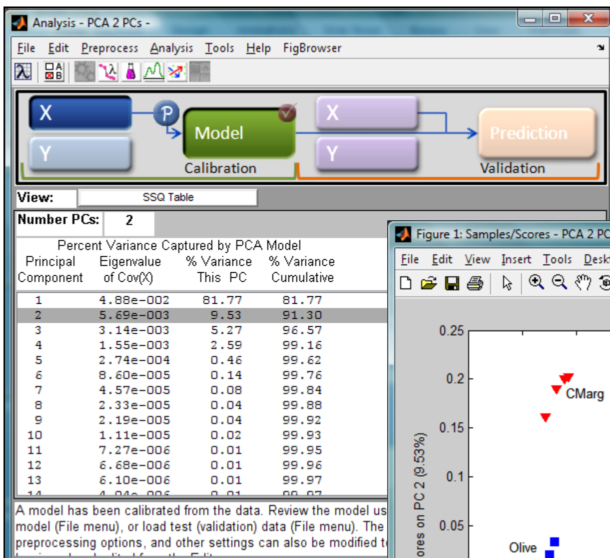
FTIR Spectra of 36 Sample Learning Set

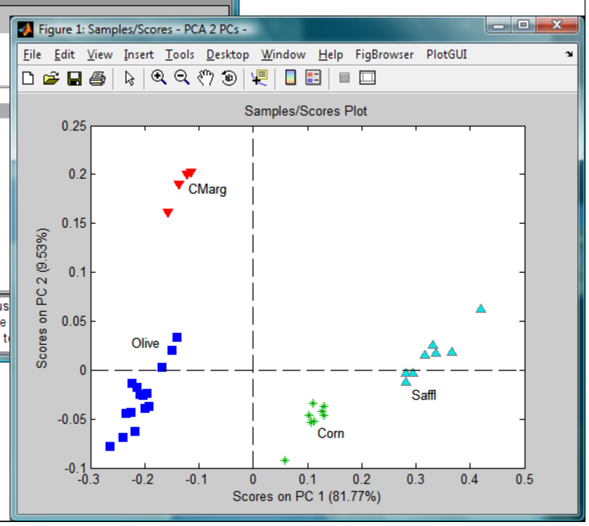


EIGENVECTOR RESEARCH INCORPORATED

173

PCA Results

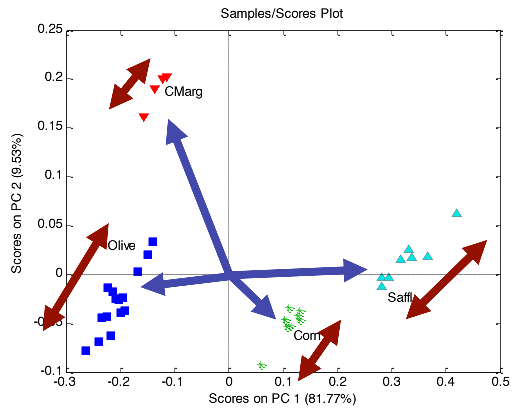




174

PCA Results

- PCA shows that the four classes in the calibration data set are separate from each other (high between class variance) but …
- have significant within class variance

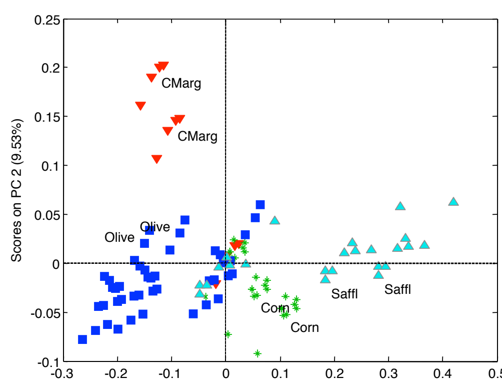


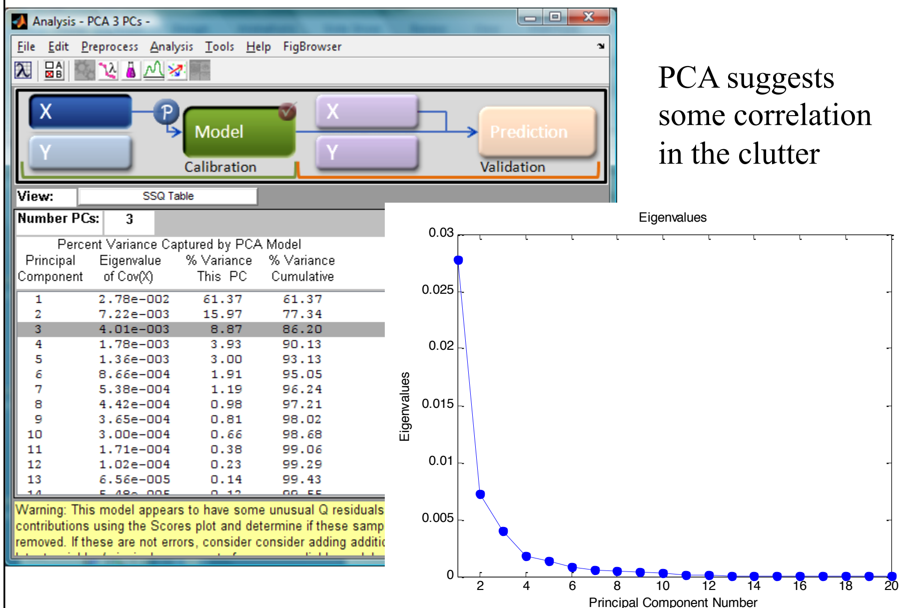
175

EIGENVECTOR RESEARCH INCORPORATED

Replicates

- Ideally, replicates would lie on top of each other.
- Variance within each class is clutter variance.
  - Is it random noise? Is the clutter correlated?
- Center each class to it's own mean and do PCA on the result.





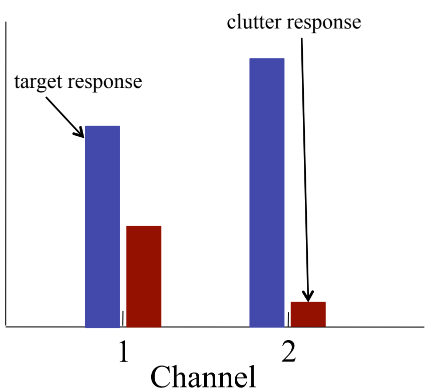
PCA suggests some correlation in the clutter

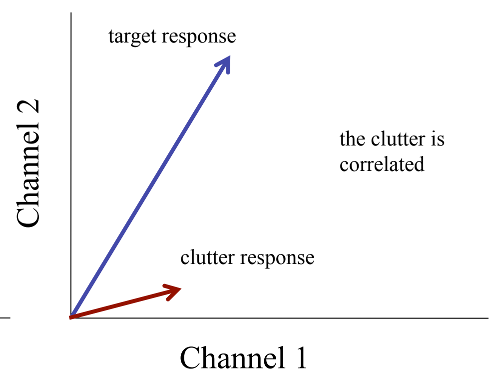
177

EIGENVECTOR RESEARCH INCORPORATED

How does the clutter affect the measurements?

- Imagine a 2-channel spectrometer





178

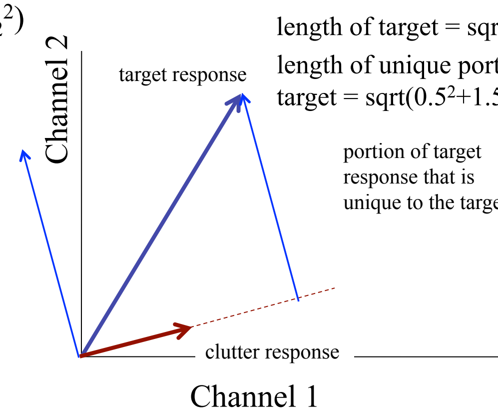
EIGENVECTOR RESEARCH INCORPORATED

How does the clutter affect the measurements?

- characterize the signal as the length of the vector $\mathrm{sqrt}(x_1^2+x_2^2)$

length of target = $\mathrm{sqrt}(1^2+2^2)=2.24$

length of unique portion of the target = $\mathrm{sqrt}(0.5^2+1.5^2)=1.58$



179

Why is clutter bad?

- The signal-to-clutter is ~proportional to the length of the unique portion of the target's response.
  - in absence of clutter it was 2.24
  - in the presence of clutter it was 1.58
- In regression, clutter-to-signal is related to the estimation error.
  - higher clutter-to-signal ➔ higher estimation error
  - in the presence of clutter the estimation error is 2.24/1.58 times the error when clutter is absent

180

Effect of Clutter

- The effect of clutter is to remove target signal
  - for olive oil example the target signal is the differences between the classes
- Instrument related clutter can be minimized by
  - good instrument design that accounts for the environment (noise+interferences) in which measurements are to be made
  - instrument standardization
    - remove drifts in offsets and gains that adds to the clutter
- Can't always be eliminated ➔ what to do?

181

EIGENVECTOR RESEARCH INCORPORATED

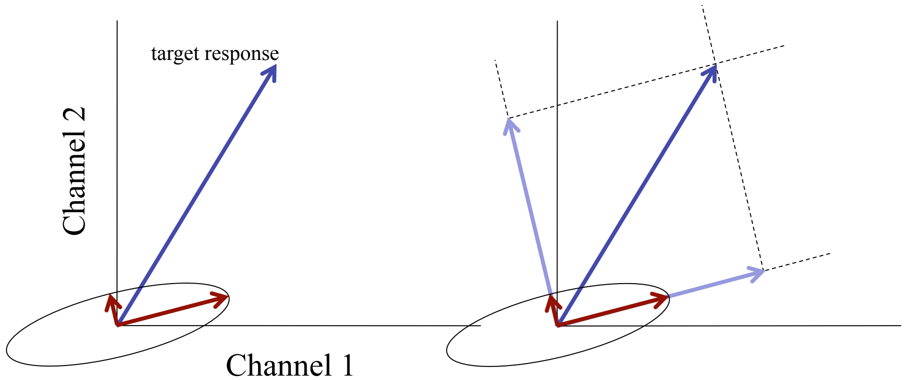
Accounting for Clutter

- One method used to account for clutter is a weighting scheme
  - similar to that used in generalized least squares (GLS)
- Autoscaling scales each variable to unit variance
- GLS weighting scales each clutter direction (as determined using PCA) to unit variance
  - directions of high clutter are deweighted
  - directions of low clutter are given more opportunity to allow signal through

182

EIGENVECTOR RESEARCH INCORPORATED

Target Projected onto Clutter

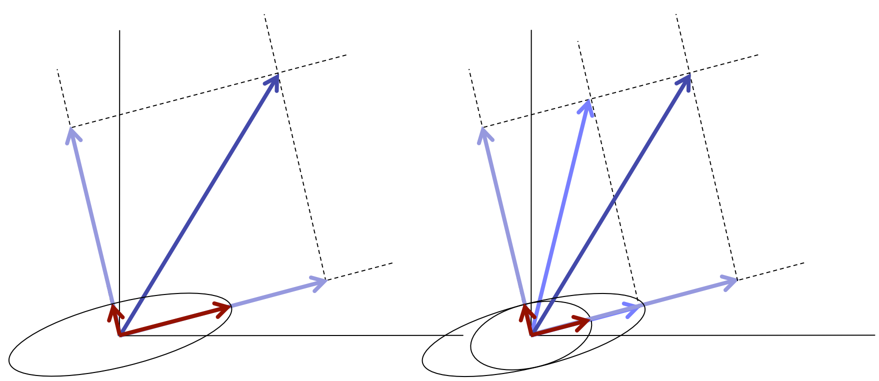


PCA of (correlated) clutter

183

EIGENVECTOR RESEARCH INCORPORATED

Scale Target by Clutter

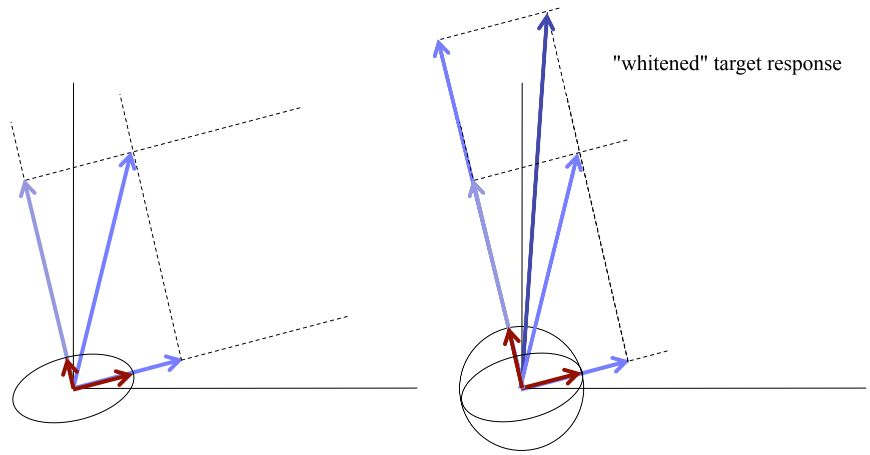


scale by first clutter PC

184

EIGENVECTOR RESEARCH INCORPORATED

Scale Target by Clutter



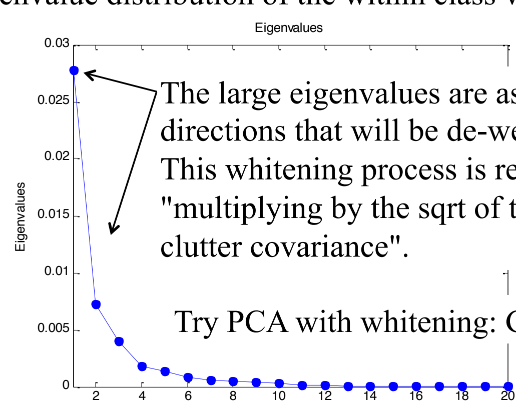
scale by first clutter PC

185

EIGENVECTOR RESEARCH INCORPORATED

Olive Oil Clutter

Eigenvalue distribution of the within class variance.



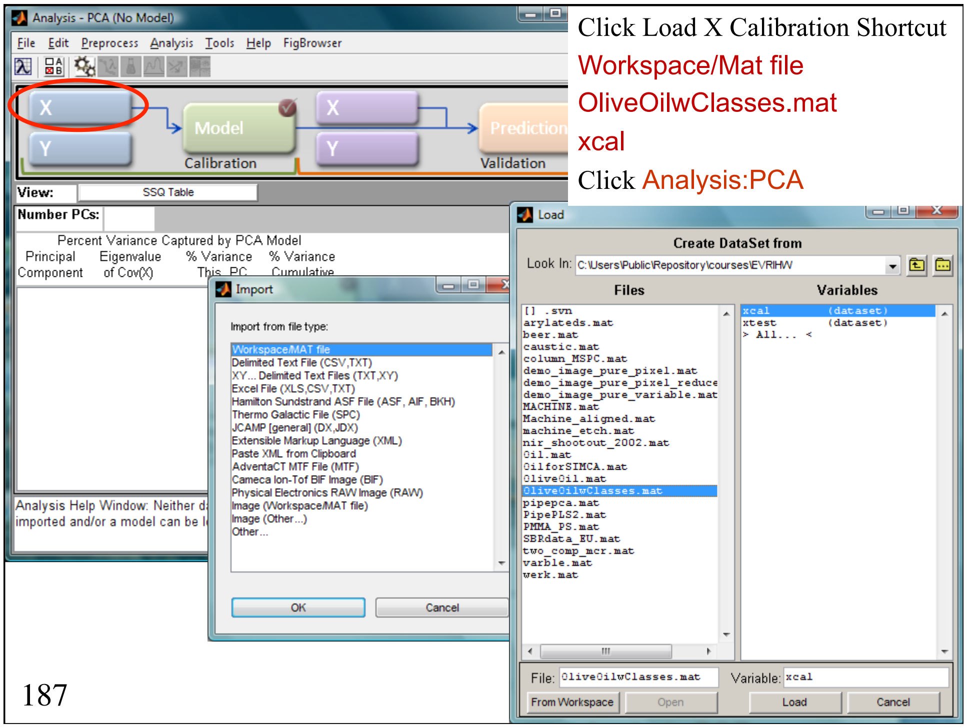
The large eigenvalues are associated with directions that will be de-weighted.
This whitening process is referred to as "multiplying by the sqrt of the inverse clutter covariance".

Try PCA with whitening: GLS weighting

186

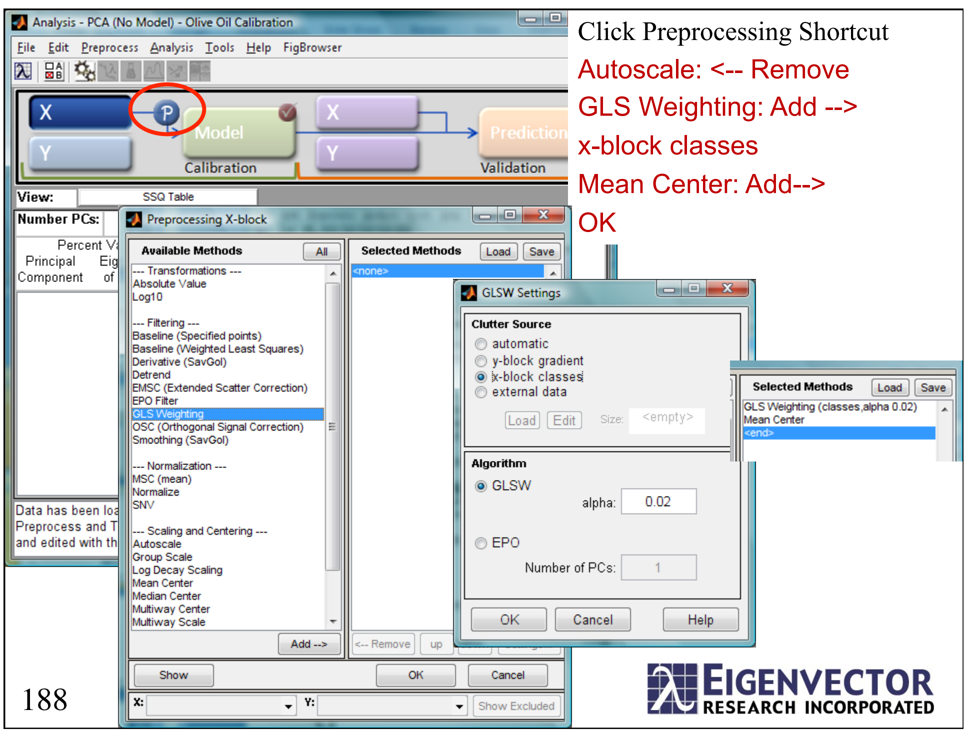
EIGENVECTOR RESEARCH INCORPORATED

Click Load X Calibration Shortcut
Workspace/Mat file
OliveOilwClasses.mat
xcal
Click Analysis:PCA



187

Click Preprocessing Shortcut
Autoscale: <-- Remove
GLS Weighting: Add -->
x-block classes
Mean Center: Add-->
OK



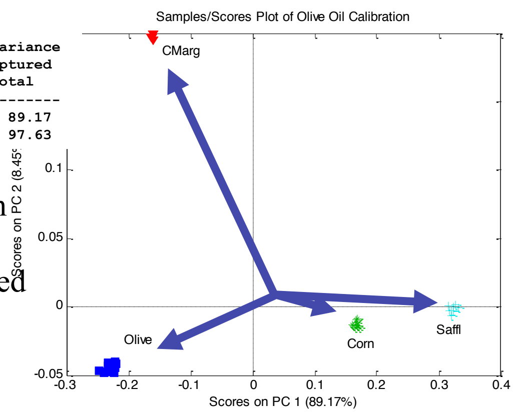
188

EIGENVECTOR RESEARCH INCORPORATED

PCA of Whitened Spectra

Percent Variance Captured by PCA Model

Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	5.63e-002	89.17	89.17
2	5.34e-003	8.45	97.63



- PCA shows the four classes in the calibration data set are further apart than non-whitened data (very high between class variance) and
- have reduced within class variance

EIGENVECTOR RESEARCH INCORPORATED

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 samples	(#37-41)
5, 10, 20, 30 & 40%		
Almond Oil	1 sample	(#42)
Peanut Oil	1 sample	(#43)
Sesame Oil	1 sample	(#44)

* New Samples not included in the calibration

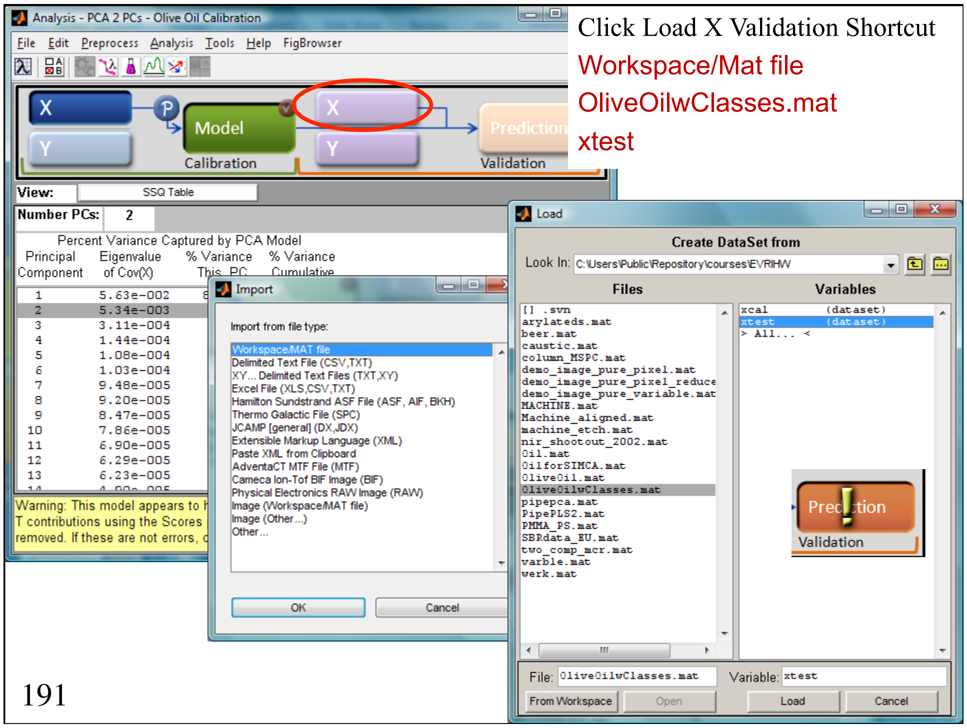
EIGENVECTOR RESEARCH INCORPORATED

Click Load X Validation Shortcut

Workspace/Mat file

OliveOilwClasses.mat

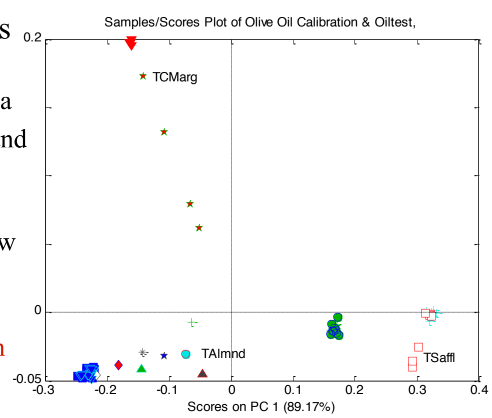
xtest



191

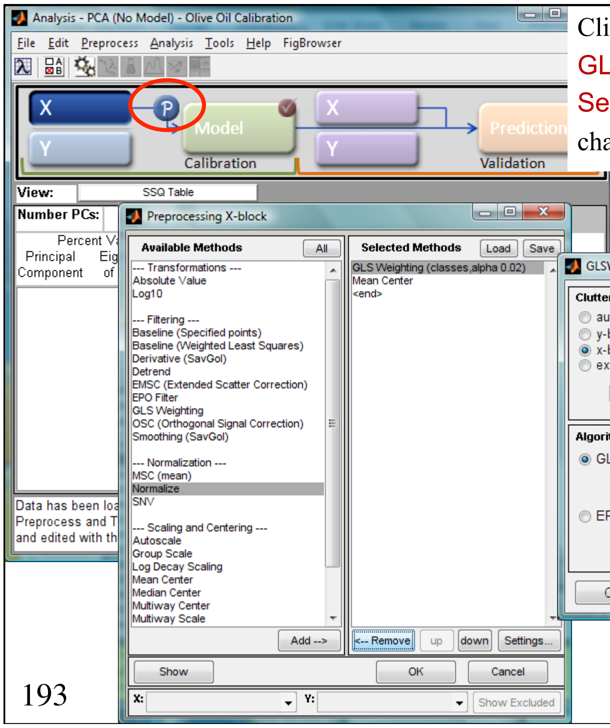
Check Test Data

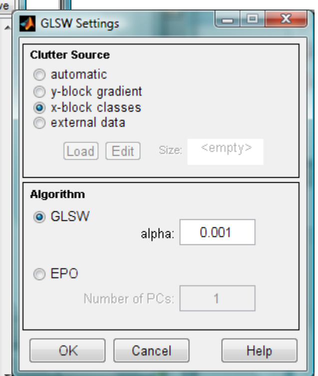
- test samples are filled symbols
  - Olive Oil and Corn appear to cluster on top of calibration data
  - This is less true for Safflower and especially Margarine, why?
  - The Margarine samples might suggest a magnitude effect. How is the effect of magnitude removed?
  - multiplicative scatter correction
  - standard normal variate
  - spectral normalization



192

EIGENVECTOR RESEARCH INCORPORATED





Click Preprocessing Shortcut

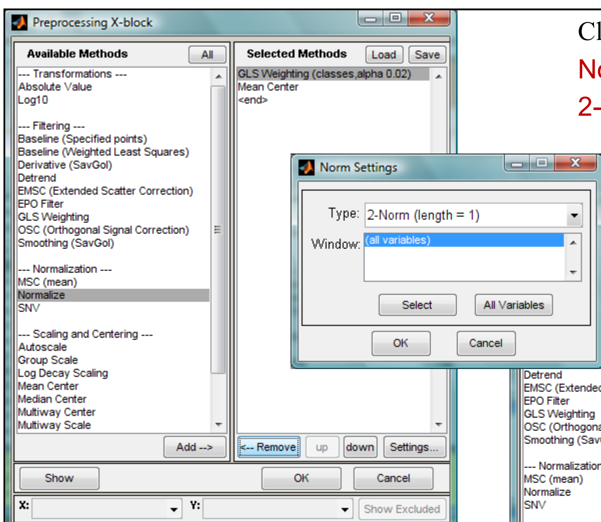
GLS Weighting (RHS)

Settings…

change alpha: 0.001

EIGENVECTOR RESEARCH INCORPORATED

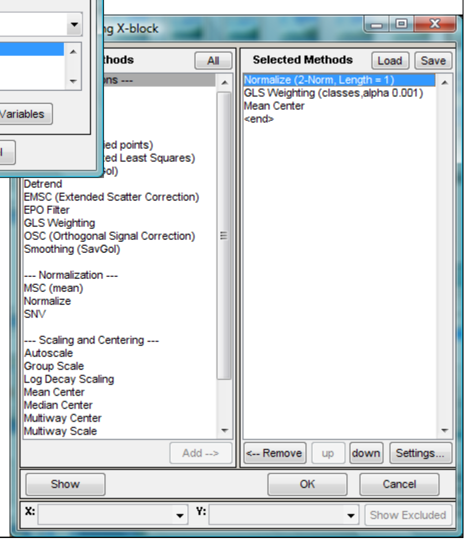
193



Click Preprocessing Shortcut

Normalize: Add -->

2-Norm



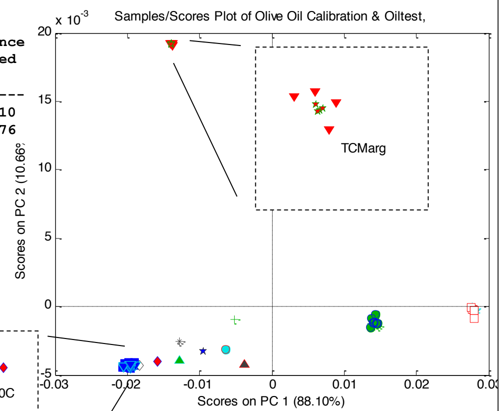
194

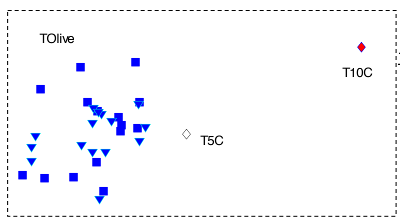
With Row Normalization

Percent Variance Captured by PCA Model

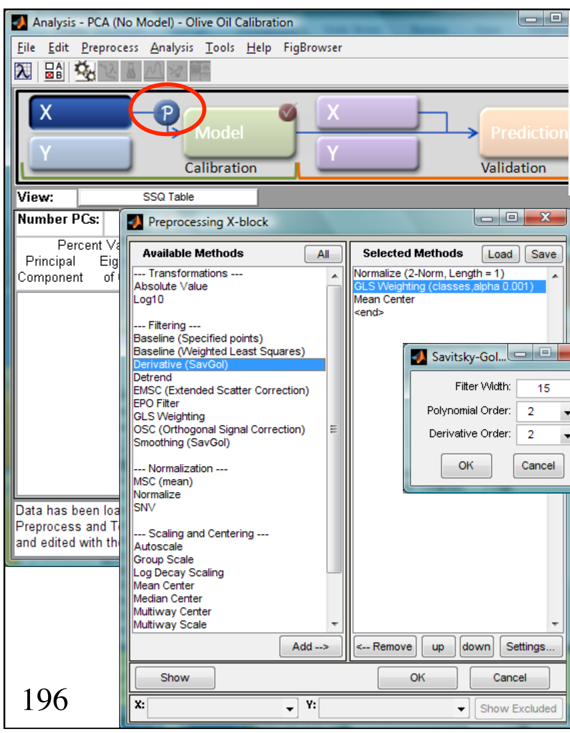
Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	4.16e-004	88.10	88.10
2	5.04e-005	10.66	98.76

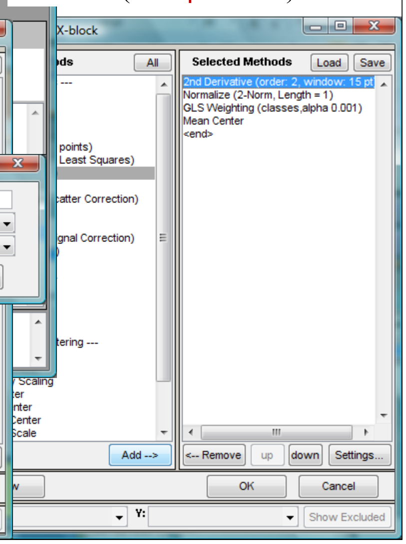
- test samples are filled symbols
  - test data are on top of calibration data
  - clusters are tight and far apart





195





Click Preprocessing Shortcut

Derivative SavGol: Add -->

Derivative Order 2

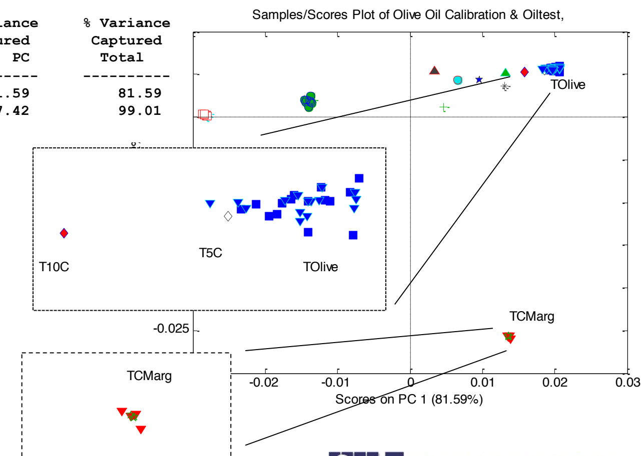
change ensure derivative is at top of list (use up / down)

196

Derivative Followed by Row Normalization

Percent Variance Captured by PCA Model

Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	4.22e-004	81.59	81.59
2	9.01e-005	17.42	99.01



197

GLS Weighting

- GLS Weighting of the spectral data accounted for some of the clutter observed in the spectra.
- The result was
  - clusters that were further apart and
  - clusters that were tighter
  - the ratio of between-class to within-class variance was increased making discrimination easier
    - clusters were so tight and far apart that confidence bounds defining each class could be wider

198

Extended Multiplicative Scatter Correction (EMSC)

- EMSC attempts to account for
  - clutter by using an extended mixture model and
  - multiplicative effects like multiplicative scatter correction (MSC)
  - The a extended mixture model is a classical least squares-like model that is used to explicitly account for clutter (a.k.a. extended least squares).

199

EIGENVECTOR RESEARCH INCORPORATED

EMSC

Provide spectra of:

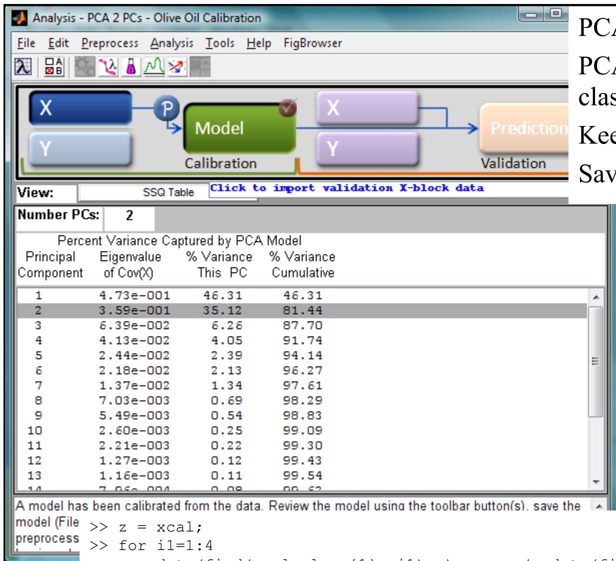
- Known ***target analytes*** **S**
- Polynomial baselines **P**
- Known ***interferences*** **Q**
  - e.g., loadings from a PCA model of clutter
  - the coefficients for each linear effect are estimated using least-squares (indicated by "hat")

$$\mathbf{s}_{2,measured} = \mathbf{s}_{ref} c_{ref} + \mathbf{S}\mathbf{c}_S + \mathbf{P}\mathbf{c}_P + \mathbf{Q}\mathbf{c}_Q \qquad \mathbf{P} = \left[\mathrm{L} \quad \boldsymbol{\upsilon}^2 \quad \boldsymbol{\upsilon} \quad \mathbf{1} \right]$$

$$\mathbf{s}_{2,corrected} = \left(\mathbf{s}_2 - \mathbf{P}\hat{\mathbf{c}}_P - \mathbf{Q}\hat{\mathbf{c}}_Q \right) \Big/ \hat{c}_{ref} \qquad \mathbf{Q} = \text{loadings}$$

200

EIGENVECTOR RESEARCH INCORPORATED



PCA of clutter.

PCA of calibration data with classes centered to class mean.

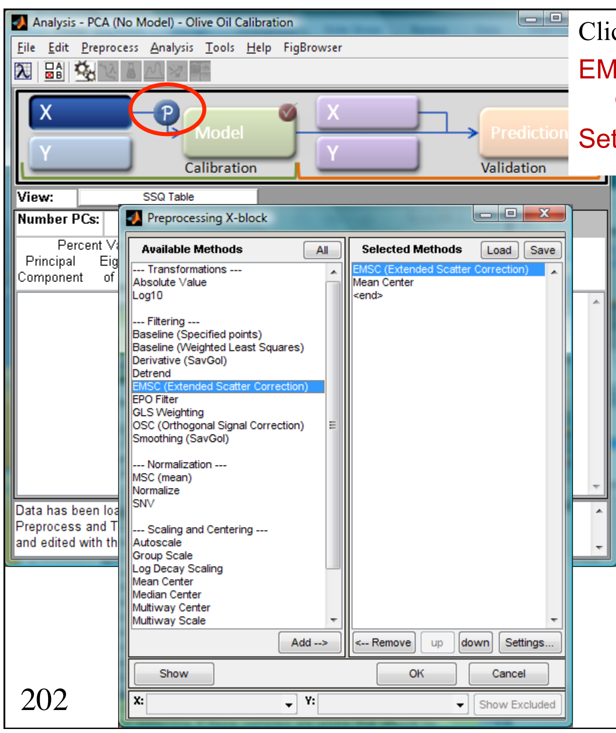
Keep 2 PCs to model the clutter.

Save model to pcam

```
>> z = xcal;
>> for i1=1:4
     z.data(find(xcal.class{1}==i1),:) = mncn(z.data(find(xcal.class{1}==i1),:));
   end
>> z.description = char(z.description,'Each class center to its own mean.');
>> p = zeros(2,518);
>> p(:,z.include{2}) = pcam.loads{2}';
```

201

EIGENVECTOR RESEARCH INCORPORATED



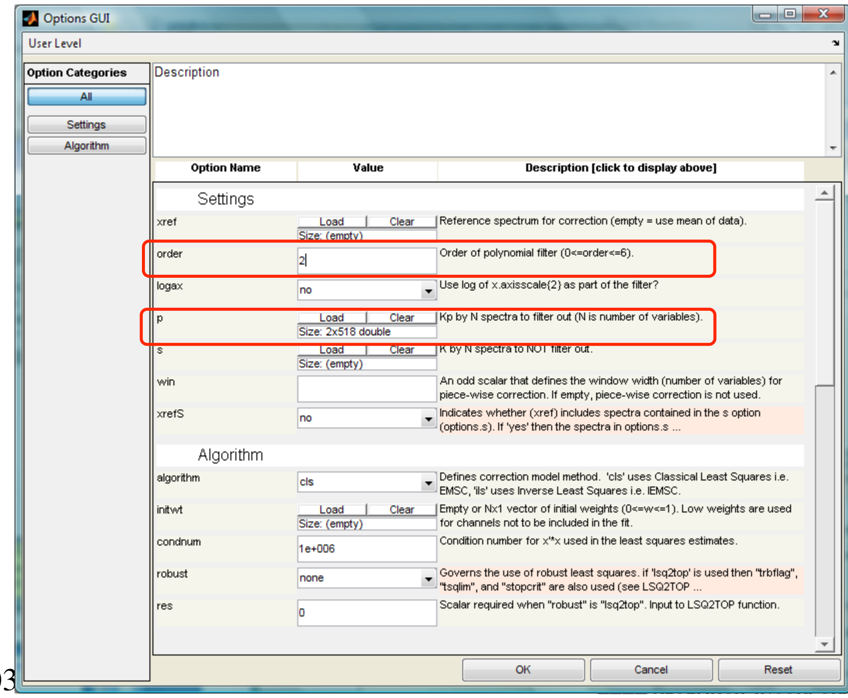
Click Preprocessing Shortcut

EMSC (Extended Scatter Correction)

Settings…

202

EIGENVECTOR RESEARCH INCORPORATED

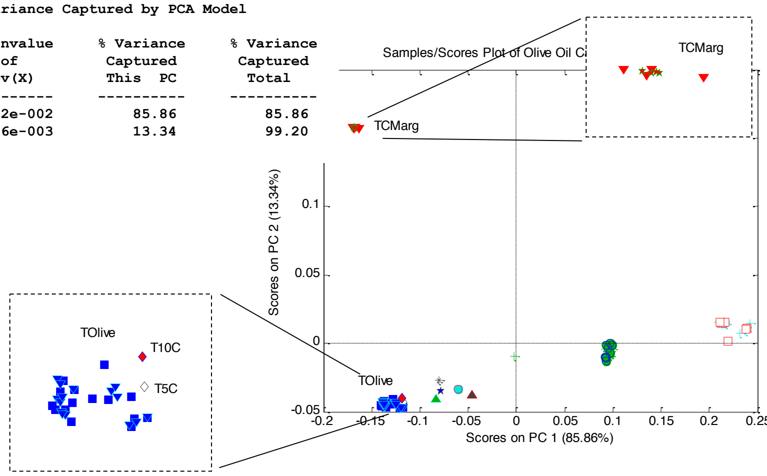


203

EMSC

Percent Variance Captured by PCA Model

Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	2.42e-002	85.86	85.86
2	3.76e-003	13.34	99.20



204

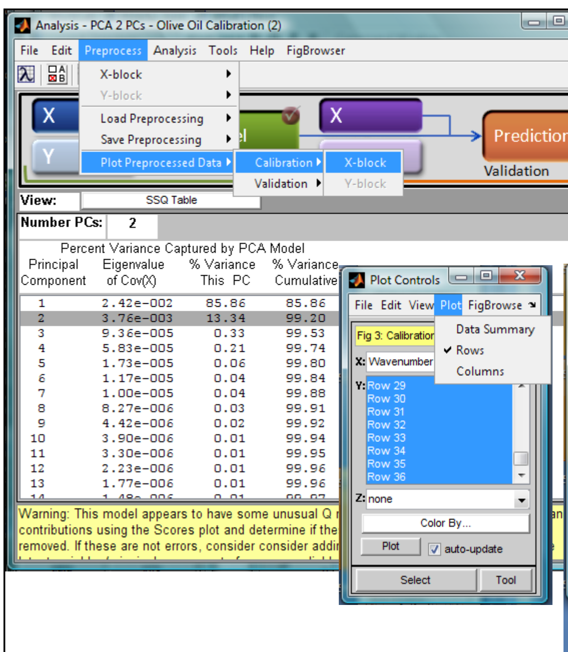
EIGENVECTOR RESEARCH INCORPORATED

EMSC Summary

- EMSC attempts to account for clutter in an explicit way
  - e.g., model clutter with basis vectors (e.g., PCA loads)
  - analyst takes control of the model
    - requires good use of measurements: clutter and target spectra
    - use what you know!
  - interpretable
    - analyst control is more daunting that using simple SavGol and MSC, but
    - the results are much more interpretable than 2nd derivative spectra

205

EIGENVECTOR RESEARCH INCORPORATED

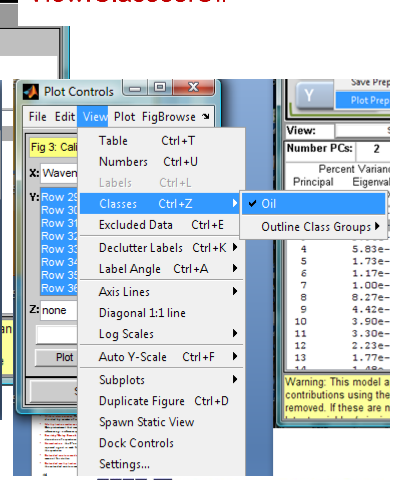


Click Preprocessing Menu

Preprocess:Plot Preprocessed Data: Calibration: X-block

Plot:Rows

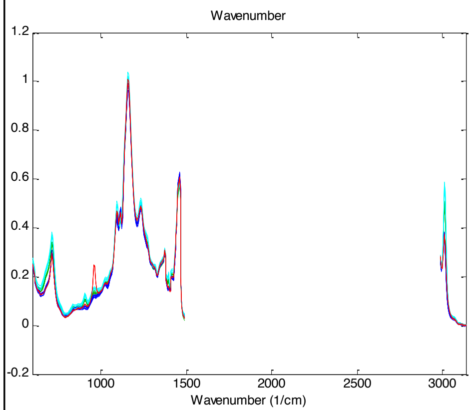
View:Classes:Oil



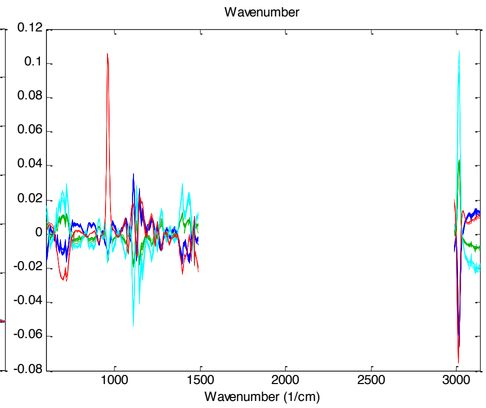
206

EIGENVECTOR RESEARCH INCORPORATED

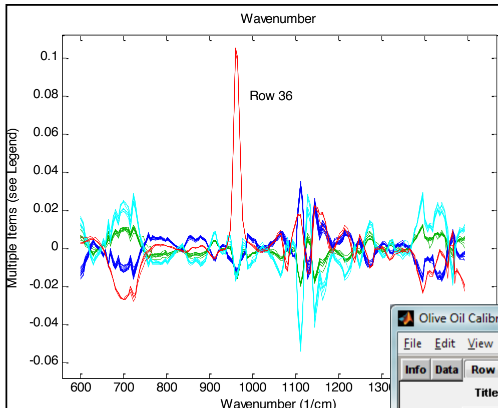
Original data

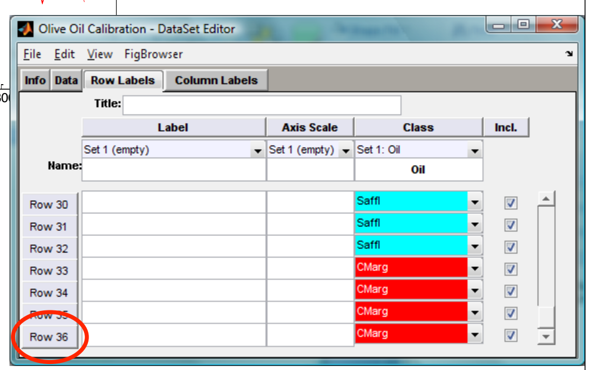


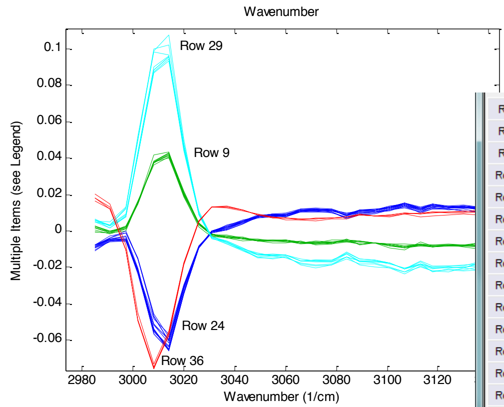
EMSC processed data, mean-centered



(colored by class)







Row 7			Corn
Row 8			Corn
Row 9			Corn
Row 10			Olive
Row 11			Olive
Row 12			Olive
Row 13			Olive
Row 14			Olive
Row 15			Olive
Row 16			Olive
Row 17			Olive
Row 18			Olive
Row 19			Olive
Row 20			Olive
Row 21			Olive
Row 22			Olive
Row 23			Olive
Row 24			Olive
Row 25			Saffl
Row 26			Saffl
Row 27			Saffl
Row 28			Saffl
Row 29			Saffl
Row 30			Saffl
Row 31			Saffl
Row 32			Saffl
Row 33			CMarg
Row 34			CMarg

209

Section Definitions

- Between class variance: The sum-of-squares of the class means centered to the global data set mean divided by number of classes.
- Within class variance: The sum-of-squares of each class centered to the class mean divided by number of samples in the class.
- Multiplicative scatter correction (MSC): (a.k.a. Multiplicative Signal Correction) Data pretreatment that removes multiplicative effects and baseline offset based on a reference *e.g.* a reference spectrum.
- Savitzky-Golay Smoothing and Differentiation: Numerical method for calculating the derivative of a spectrum that uses windowed polynomials.
- Normalization: the 2-norm divides a spectrum by the square root of the sum-of-squared signal in each frequency channel. This removes magnitude information from the spectrum.
- Extended mixture model: a classical least squares-like model that is used to explicitly account for clutter.
- Extended multiplicative scatter correction (EMSC): a model that combines MSC and the extended mixture model to explicitly account for clutter.

210

Outline

- Introduction
- Advanced Preprocessing
  - Clutter and characterizing clutter
  - Generalized least squares weighting
  - Extended multiplicative scatter correction
  - Interval PLS (iPLS)
  - Model Robustness
- Multivariate image analysis
- Multi-way Analysis
- Summary

EIGENVECTOR RESEARCH INCORPORATED

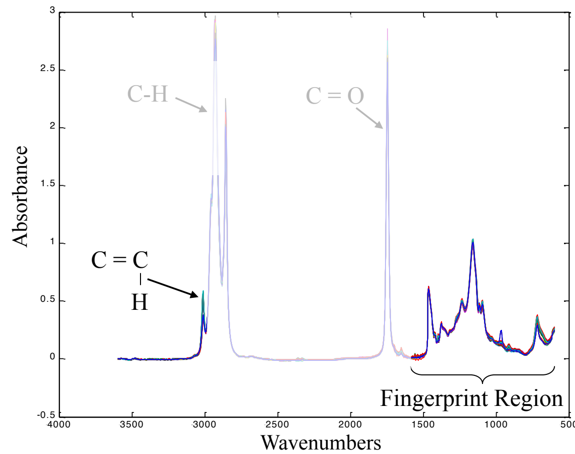
Why Variable Selection?

- Improvement of the model
  - Remove irrelevant, unreliable or noisy variables (clutter)
  - Improve predictions
  - Improve statistical properties
- Interpretation
  - Obtain a model that is easier to understand
- Costs
  - Use fewer measurements to replace expensive or time-consuming one
- Development of fast instruments/routines for on-line control
  - Find wavelength ranges for a filter-based instrument

212

EIGENVECTOR RESEARCH INCORPORATED

Already done some based on *a posteriori* knowledge…



213

EIGENVECTOR RESEARCH INCORPORATED

Variable Selection Methods

- ***a priori***
  - Choose measurements
- ***a posteriori***
  - Use chemical/physical insight
- **Model based**
  - Look at loadings
- **"Random based"**
  - Genetic algorithms
  - Simulated annealing
- **“Spectral”**
  - i-PLS
  - fullsearch
- **Classical**
  - Forward, backward selection
  - Best subset selection
  - Significance tests
  - Significance based on Jack-knife
  - GOLPE
- **Other**
  - Pure variables
  - Principal variables
  - Iterative weighting with regression vector
  - …

(see the Variable Selection Course at EigenU)

214

EIGENVECTOR RESEARCH INCORPORATED

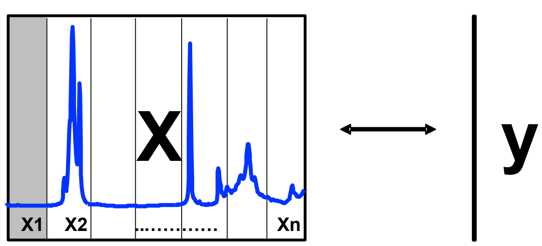
Variable Selection Methods

- How to choose which method?!?
- Different methods work in different situations
- Interval-PLS is a good “example” method to understand the considerations of variable selection. Simple to implement and use.

215

i-PLS Theory

iPLS: Interval PLS
Build local models using “intervals” of X-block variables. Very intuitive and useful approach that can be easily combined with variable selection.



L. Nørgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, S. B. Engelsen. Interval partial least-squares regression (iPLS). *Appl.Spectrosc.* 54 (3):413-419, 2000.

216

Fit Criteria

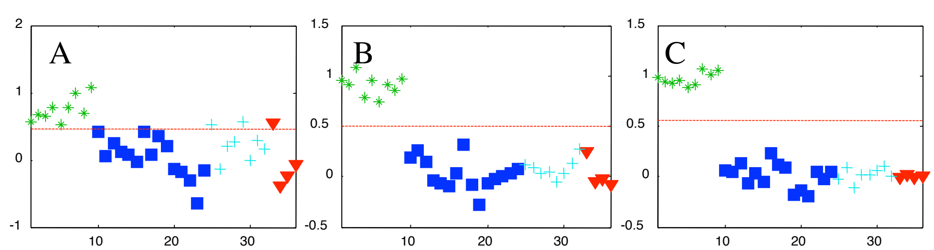
- RMSECV is used to determine “best” interval
- For Olive Oil data: perform discriminant analysis using PLSDA (use logical y-block with PLS to separate classes).
- Note: Always validate afterwards! Variable selection methods have a tendency to give over-optimistic RMSE results.

217

EIGENVECTOR RESEARCH INCORPORATED

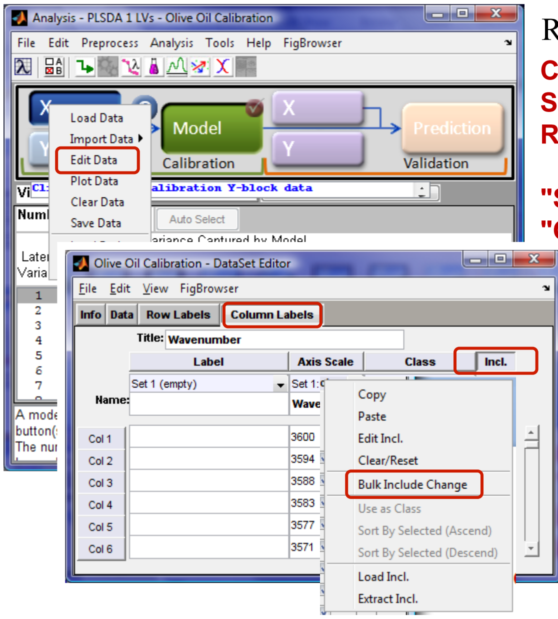
RMSE vs. Misclassification Rate in PLSDA

- RMSE shows deviation from predicting 0 or 1.
- Misclassification Rate shows prediction on "wrong side" of decision line.
- RMSE: A>B>C   Misclassification: A>B=C



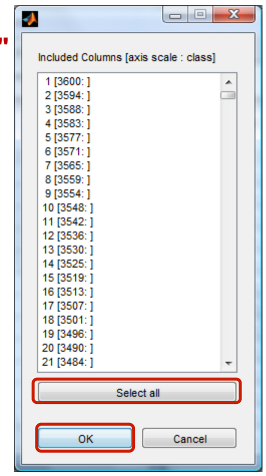
218

EIGENVECTOR RESEARCH INCORPORATED



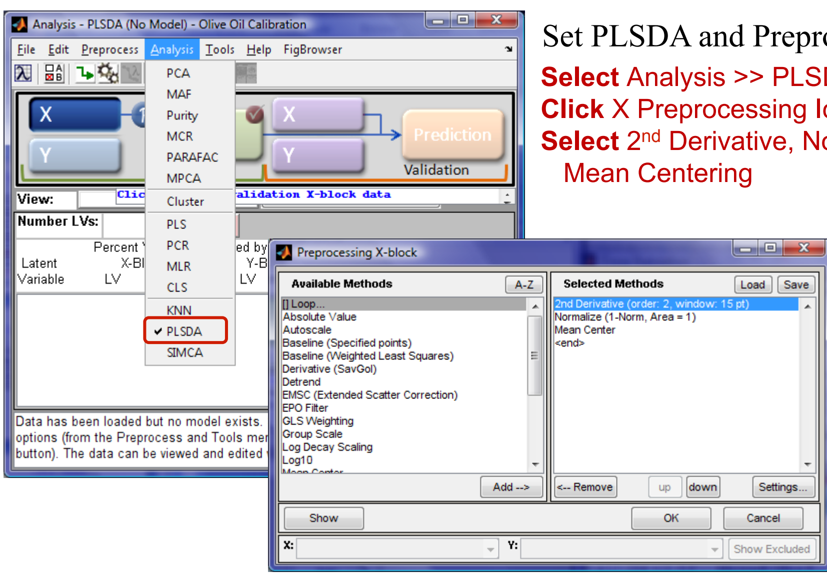
Re-Include all variables

Click "X" >> Edit Data
Select Column Labels tab
Right Click Incl. >> Bulk Include Change
"Select all"
"OK"



219

EIGENVECTOR RESEARCH INCORPORATED

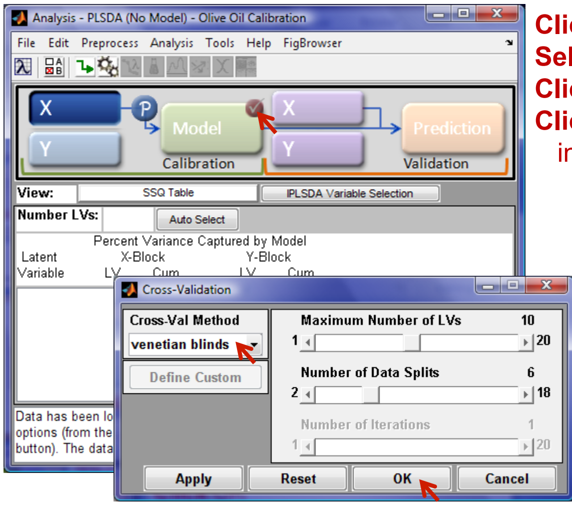


Set PLSDA and Preprocessing

Select Analysis >> PLSDA
Click X Preprocessing Icon
Select 2nd Derivative, Normalize, Mean Centering

220

EIGENVECTOR RESEARCH INCORPORATED

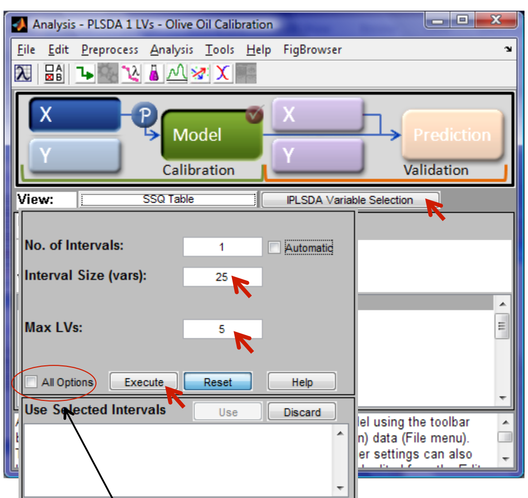


Click Cross Validation Icon
Select "Venetian Blinds"
Click OK
Click "Model" (to build model initial PLSDA model)

Normally, should now review scores for outliers!

221

EIGENVECTOR RESEARCH INCORPORATED



Click "IPLSDA Variable Selection" button
Enter 25 for Interval Size
Enter 5 for max LVs
Click "Execute"

Interval Size defines how many variables to group together.

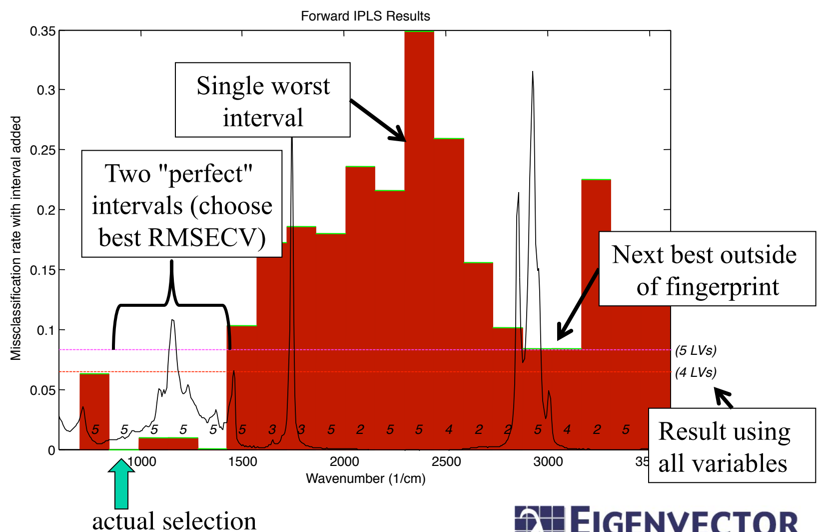
Spectral data: use size>1
Non-contiguous data: use size=1 (single variable intervals)

provides access to other i-PLS settings

222

EIGENVECTOR RESEARCH INCORPORATED

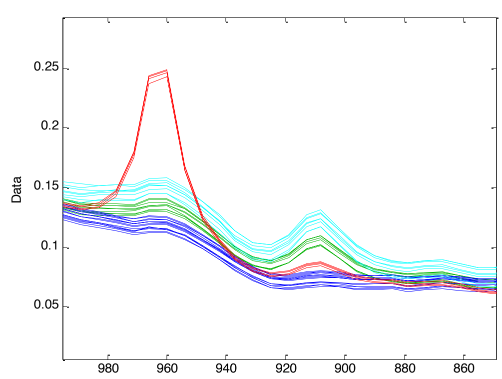
First Interval Result



223

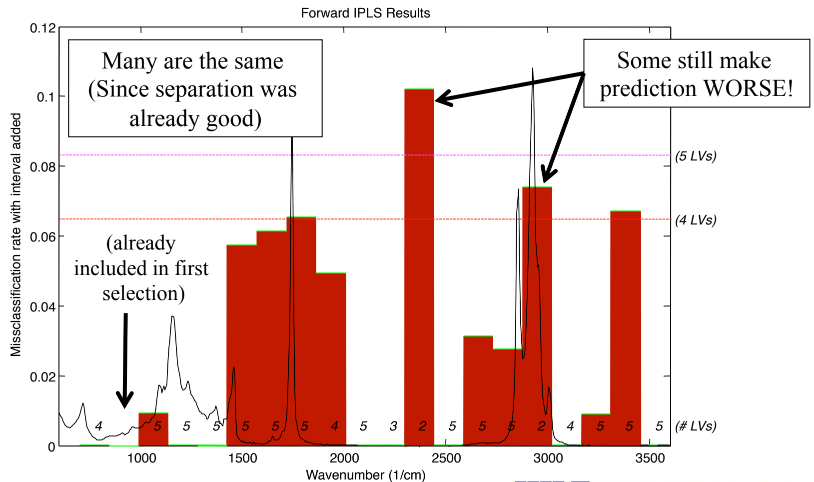
Why the Low-Frequency Intervals? Information Content

- Best intervals contain useful signal from all four classes.
- Original data shown; Preprocessing only makes this separation better!



224

Repeat Execution and "Add To" Existing



225

EIGENVECTOR RESEARCH INCORPORATED

Number of Intervals

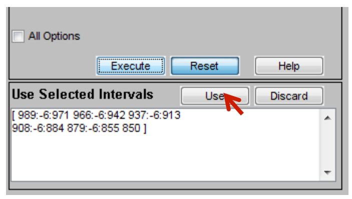
- Can choose a pre-set number of intervals to find
- Can also use "Automatic" to continue selecting intervals until RMSECV/misclassification does not improve
- This is **not** the same as exhaustive combinatorial search (fullsearch). It is sequential (choose one, "lock" it in, choose a second, "lock" it in…)
- For very complex data, may not give actual "best" windows, but probably not a bad one.

226

EIGENVECTOR RESEARCH INCORPORATED

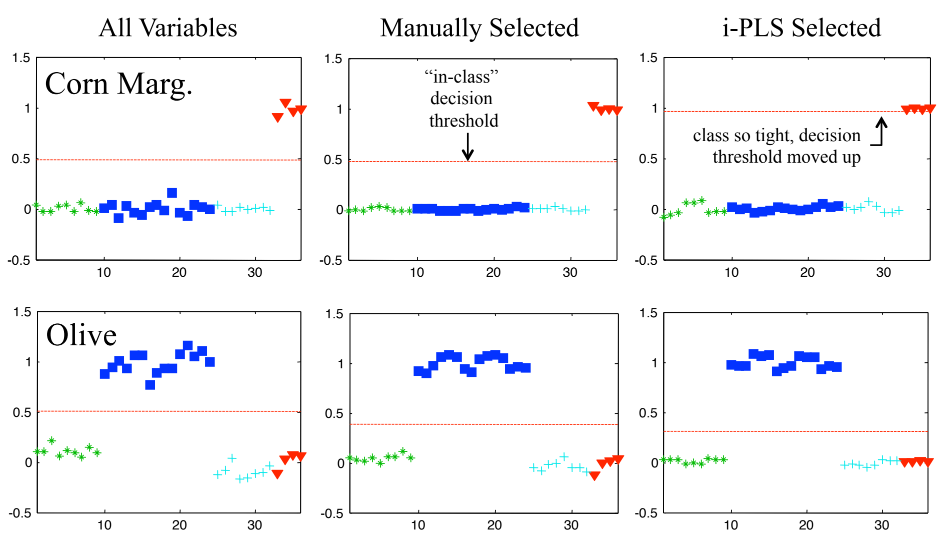
What is the Result?

- For PLSDA, lower RMSECV should indicate better class separation in predicted Y values
- Selecting additional intervals gives little improvement in RMSECV (on this data)
- Use ONLY first selected interval and build new model…



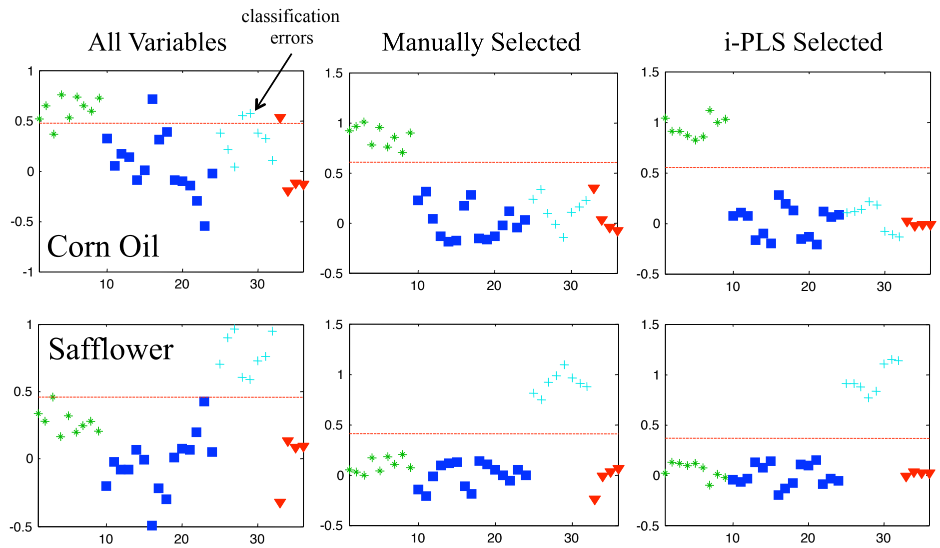
227

Cross-validated Predictions



228 Mild improvement... (note: LVs = 5, 5, 3)

Cross-validated Predictions



229 i-PLS: ~tighter clusters than manual

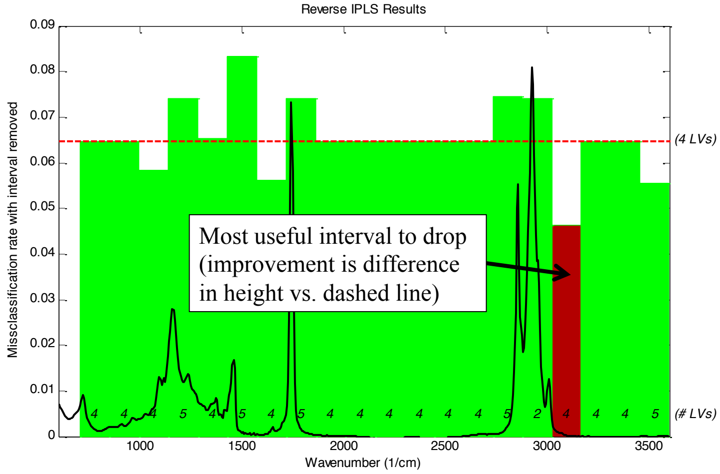
Reverse i-PLS

- **Principle**
  - Make full model and select the variable contributing the least to the fit (exclude regions which contain **more clutter**)
  - Repeat as desired/needed
- **Good**
  - Takes interactions into account
  - Reasonably fast
- **Bad**
  - ”Random” removal for large data sets
  - Often works bad for many irrelevant variables

(see “All options” checkbox on i-PLS controls)

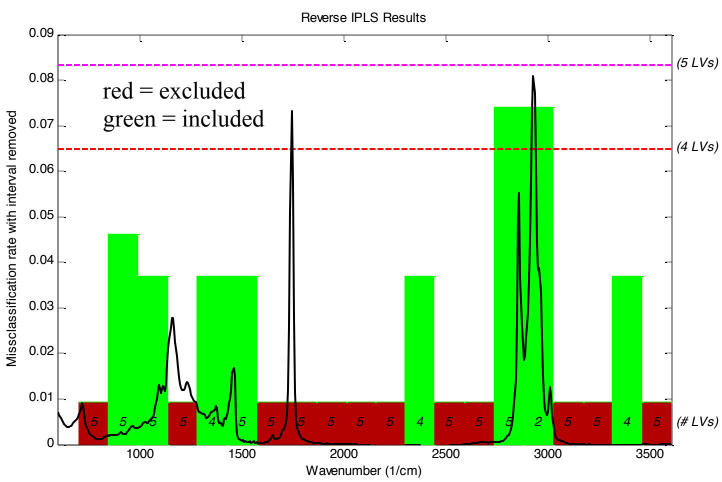
230

Reverse i-PLS



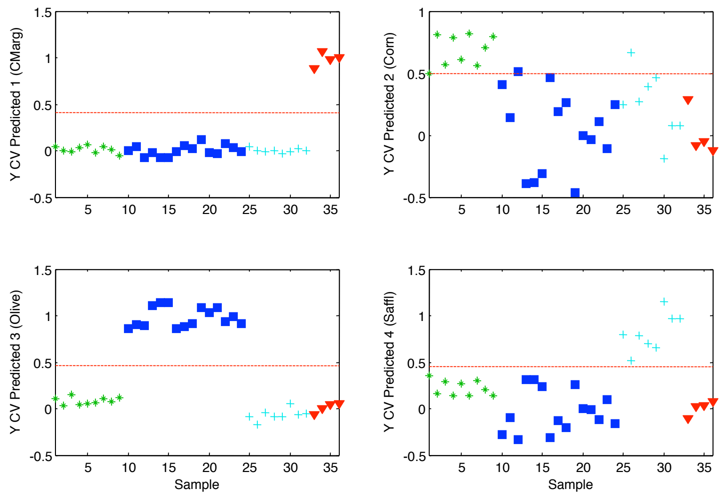
231

Automatic Reverse i-PLS



232

Cross-validated Predictions after Automatic Reverse i-PLS



8 intervals retained. Class 2 and 4 not as good as forward i-PLS!

233

EIGENVECTOR RESEARCH INCORPORATED

What is Model Robustness?

- When developing calibration models focus is generally on improving prediction error
- Models often developed with small amount of data taken over relatively short time
- Prediction errors over long term often dominated by artifacts not represented in calibration data
  - Changes in spectrometer / sensor
  - Changes in sample

234

EIGENVECTOR RESEARCH INCORPORATED

Typical Changes in the System

- Sample
  - New analyte(s)
  - Changes in physical properties (e.g. scattering)
  - Temperature
  - Pressure
- Instrument (spectrometers)
  - Wavelength/Frequency registration shift
  - Stray light
  - Resolution
  - Noise

235

What Constitutes a Good Model?

- Acceptable prediction error (not necessarily the best achievable)
- Longevity, i.e. ***robustness*** to minor changes

- Once you have built a model, you should exercise it with expected changes
- Use real data or simulate typical instrumental and multivariate errors

236

Robustness Testing

- Develop model with desired preprocessing, #LVs, etc.
- "Perturb" test data set
- Apply calibration model to "perturbed" data
- Look at prediction error as function of perturbations
- Test and compare multiple models

237

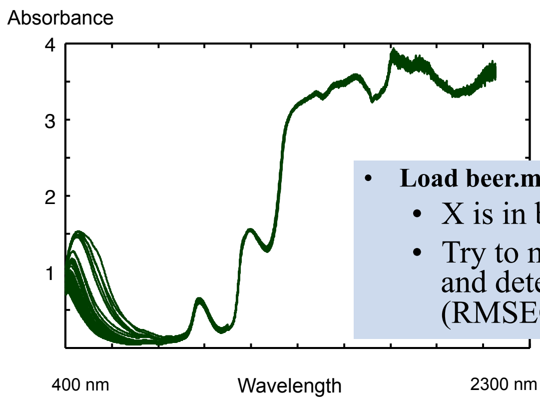
Perturbations

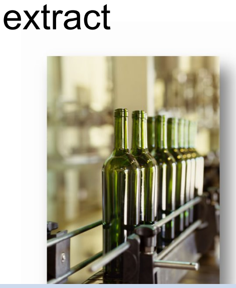
- New analyte – add Gaussian peak of variable width across wavelength range
- Wavelength registration shift – shift spectra left-right as well as expand and contract
- Baseline shift – change offset and slope
- Stray light – add fraction of signal before log transform
- Temperature – decrease resolution and vary path length
- Noise variation – add noise with varying bandwidth

238

VIS/NIR spectra of 61 beers

Purpose: prediction of real extract





- **Load beer.mat**
  - X is in beer and Y in extract
  - Try to make a nice PLS model and determine quality (RMSEC, RMSECV)

239

EIGENVECTOR RESEARCH INCORPORATED

Exercise Data

Determination of the amount of extract from NIR spectra of beers.

Dispersive visual & near-infrared data collected (at 25 C) NIRSystems Inc. (Model 6500) spectrophotometer. Split detector system – silicon detector 400-1100 nm & (PbS) detector 1100-2500 nm.

VIS-NIR transmission recorded directly on undiluted degassed beer in 30 mm quartz cell. Spectral data collected at 2 nm intervals 400-2250 nm & converted to absorbance units.

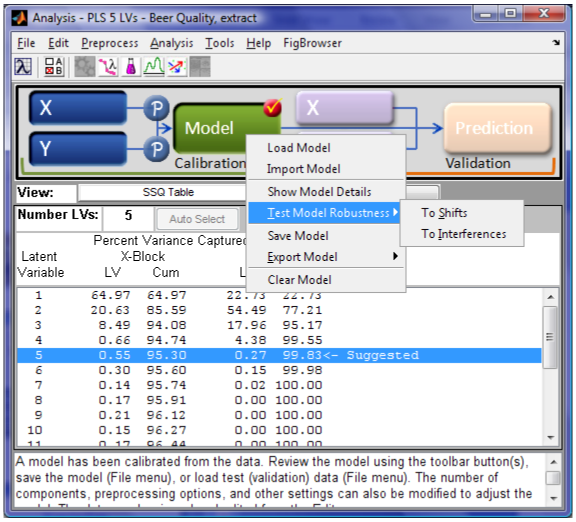
Original *extract* concentration is a quality parameter in the brewing industry, indicating the substrate potential for the yeast to ferment alcohol and serving as a taxation parameter. Original extract concentration determined by Carlsberg A/S in the range of 4.23-18.76% plato.

Data sorted by extract value, and a model independent test set was constructed by selecting every third sample of this full data set. There are thus two data sets: one for calibration (40 samples) and one for independent estimation of prediction error (20 samples).

240

EIGENVECTOR RESEARCH INCORPORATED

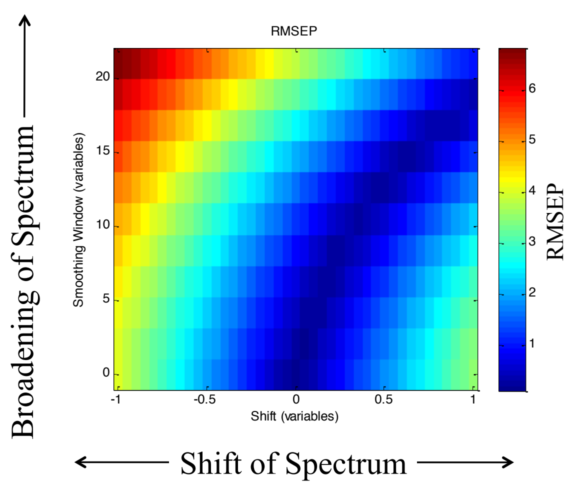
Test Robustness



Two built-in tests:
- Shift / Resolution Test
- Interference Test

241

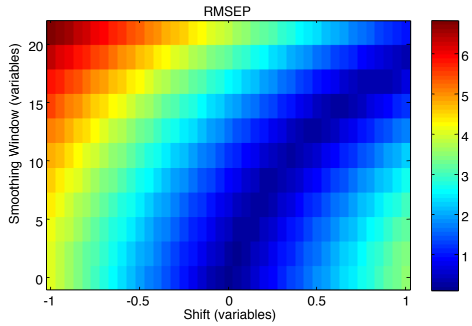
Shift / Resolution Test



Model
Normalization
+ Autoscaling
RMSEC: 0.10
RMSECV: 0.34

Hint: use “caxis” command to get color range and “caxis([min max])” command to set color range

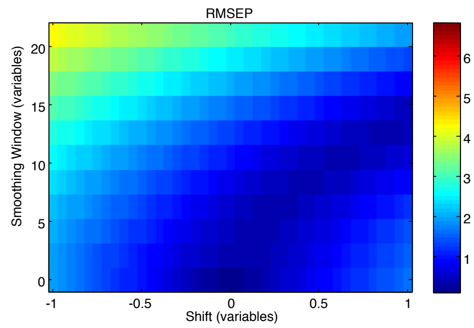
242



Model A

Normalization
+ Autoscaling
5 Latent Variables
RMSEC: 0.10
RMSECV: **0.34**

Model B starts worse, but is less influenced by shift!

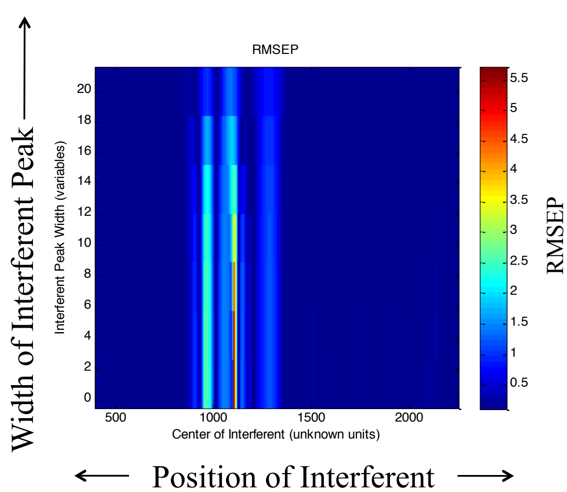


Model B
Detrend
+ Normalization
+ Autoscaling
4 Latent Variables
RMSEC: 0.14
RMSECV: **0.57**

243

EIGENVECTOR RESEARCH INCORPORATED

Interferent Test

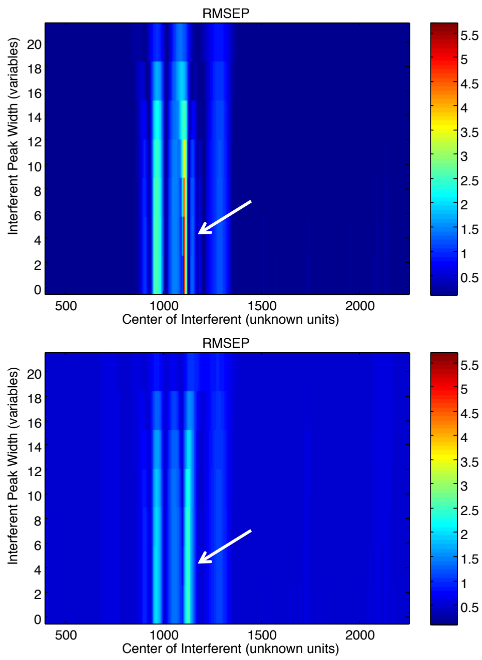


Model

Normalization
+ Autoscaling
5 Latent Variables
RMSEC: 0.10
RMSECV: **0.34**

244

EIGENVECTOR RESEARCH INCORPORATED



<u>Model A</u>

Normalization
+ Autoscaling
5 Latent Variables
RMSEC: 0.10
RMSECV: **0.34**

Model B starts worse, but is less influenced by interferent

<u>Model B</u>

Normalization
+ Autoscaling
3 Latent Variables
RMSEC: 0.53
RMSECV: **0.70**

245

EIGENVECTOR RESEARCH INCORPORATED

Outline

- Introduction
- Advanced Preprocessing
- Multivariate image analysis
- Multi-way Analysis
- Summary

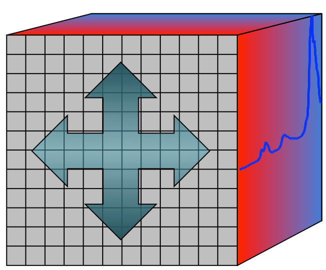
EIGENVECTOR RESEARCH INCORPORATED

Multivariate Images

A data array of *dimension three* (or more) where the first two dimensions are *spatial* and the last dimension(s) is a function of another variable.

247

Multivariate Images



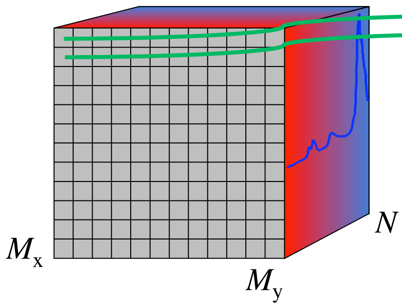
Spectral Information between channels (chemical information)

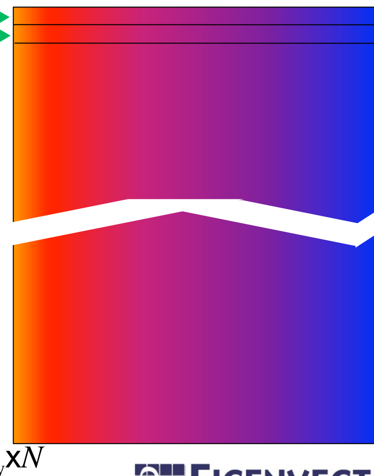
Spatial Information between pixels

Spatial distribution of chemical analytes, physical features, and other properties

248

Reshaping Images for Analysis



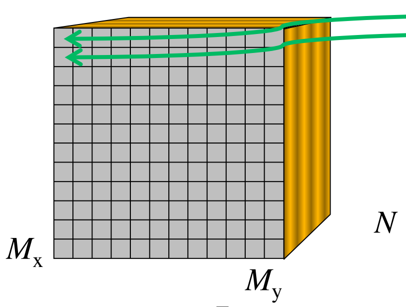


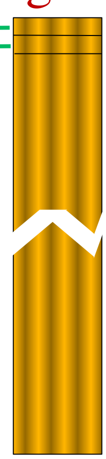
Reshape $\underline{\mathbf{X}}$ (M_xxM_yxN) to $\mathbf{X}$ (M_xM_yxN) matrix and subject to two-way analysis e.g, PCA

EIGENVECTOR RESEARCH INCORPORATED

249

PCA for Multivariate Image Analysis





PCA is an example of MIA. *Techniques and Applications of Hyperspectral Image Analysis*, Grahn, H. F.; Geladi, P., Eds. John Wiley & Sons: West Sussex, England (2007)

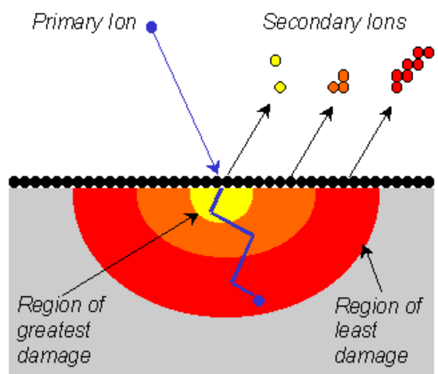
Model $\mathbf{X} = \mathbf{TP}^T\mathbf{+E}$
Reshape $\mathbf{T}$ (M_xM_yxK) to $\underline{\mathbf{T}}$ (M_xxM_yxK) and examine the scores images

EIGENVECTOR RESEARCH INCORPORATED

250

Example: TOF-SIMS

- Time-of-Flight Secondary-Ion-Mass Spectrometry
  - common surface analysis technique
  - mass spectrum generated for each pixel



Thanks to
Physical Electronics!

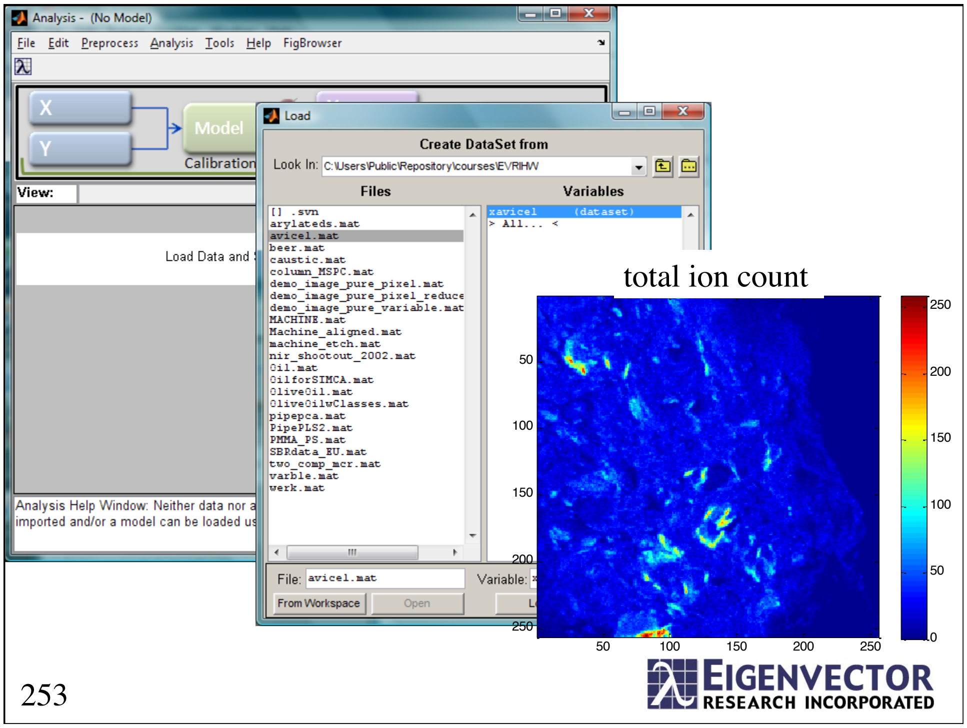
251

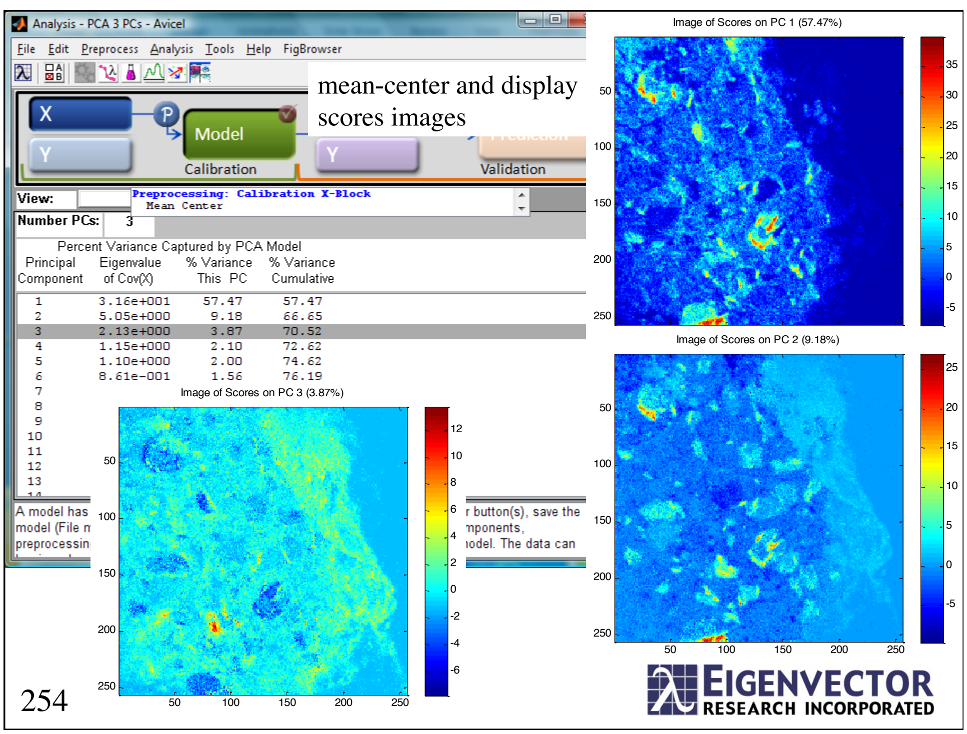
TOF-SIMS of Time Release Drug Delivery System

- Multi-layer drug beads serve as controlled release system
- TOF-SIMS of cross section of bead
- Evaluate the integrity of the layers and distribution of ingredients

A.M. Belu, M.C. Davies, J.M. Newton and N. Patel, "TOF-SIMS Characterization and Imaging of Controlled-Release Drug Delivery Systems," *Anal. Chem.*, **72**(22), 5625–5638 (2000).
Gallagher, N.B., Shaver, J.M., Martin, E.B., Morris, J., Wise, B.M. and Windig, W., "Curve resolution for images with applications to TOF-SIMS and Raman", *Chemometr. Intell. Lab.*, **73**(1), 105–117 (2003).

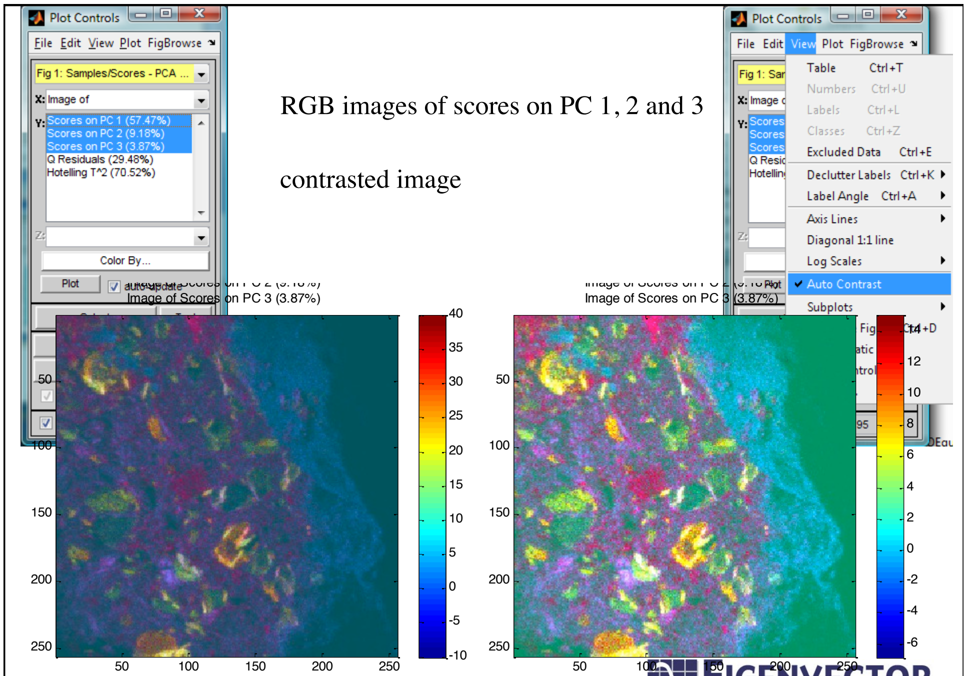
252


Analysis - (No Model)
File Edit Preprocess Analysis Tools Help FigBrowser
Model
Calibration
View:
Load
Create DataSet from
Look In: C:\Users\Public\Repository\courses\EVRIHW
Files
Variables
File: avicel.mat
Variable:
From Workspace
Open
total ion count
253
EIGENVECTOR RESEARCH INCORPORATED


Analysis - PCA 3 PCs - Avicel
File Edit Preprocess Analysis Tools Help FigBrowser
mean-center and display scores images
Model
Calibration
Validation
Preprocessing: Calibration X-Block
Mean Center
View:
Number PCs: 3
Percent Variance Captured by PCA Model
Principal Component
Eigenvalue of Cov(X)
% Variance This PC
% Variance Cumulative
1 3.16e+001 57.47 57.47
2 5.05e+000 9.18 66.65
3 2.13e+000 3.87 70.52
4 1.15e+000 2.10 72.62
5 1.10e+000 2.00 74.62
6 8.61e-001 1.56 76.19
Image of Scores on PC 1 (57.47%)
Image of Scores on PC 2 (9.18%)
Image of Scores on PC 3 (3.87%)
254
EIGENVECTOR RESEARCH INCORPORATED

RGB images of scores on PC 1, 2 and 3

contrasted image



255

Minimum Noise Factors (MNF)

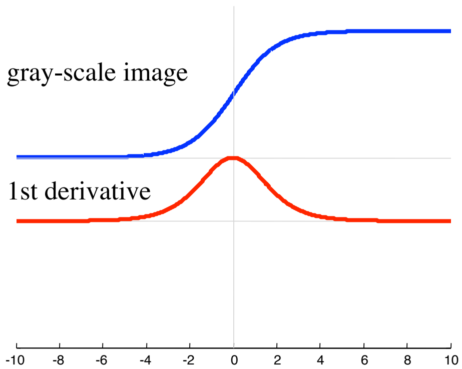
- MNF attempts find directions in the data that maximize the signal-to-clutter.

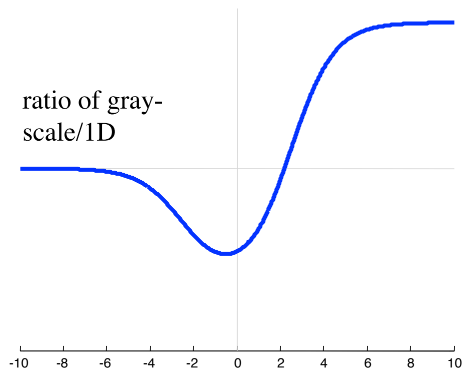
$$\max_{\mathbf{v}_i \neq 0} \left(\frac{\mathbf{v}_i^T \boldsymbol{\Sigma}_X \mathbf{v}_i}{\mathbf{v}_i^T \boldsymbol{\Sigma}_C \mathbf{v}_i} \right) \quad \text{the objective function}$$

- Result is a PCA-like eigenvector problem
- In maximum autocorrelation factors (MAF) clutter is the first difference image (difference between near-by pixels)

256

MAF





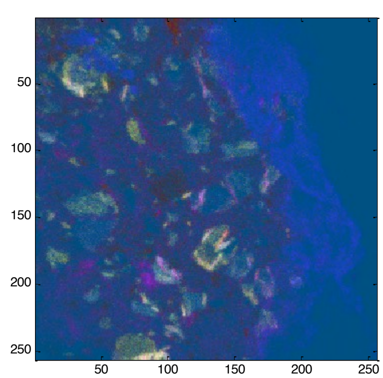
MAF finds locations in the image where the ratio of gray-scale to first derivative is a maximum

257

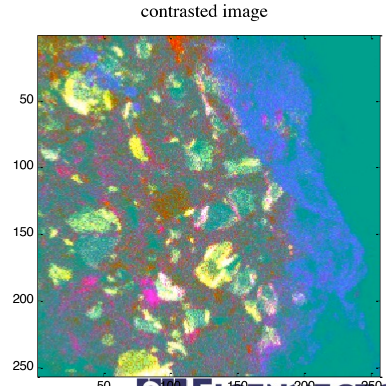
EIGENVECTOR RESEARCH INCORPORATED

MAF Results

RGB images of MAF scores on 1, 2 and 3



contrasted image

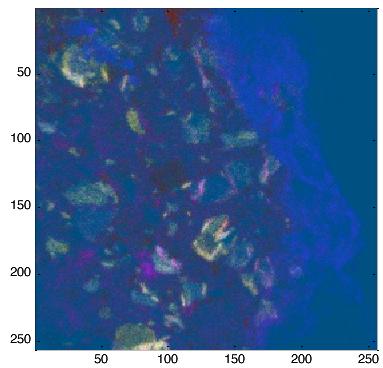


258

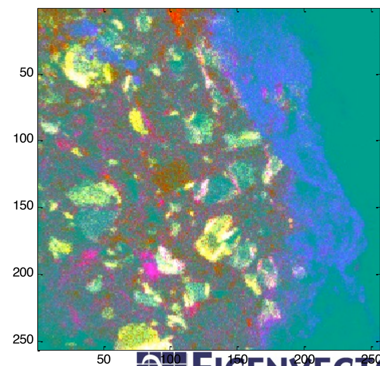
EIGENVECTOR RESEARCH INCORPORATED

PCA w/ GLS Weighting for ~MAF

RGB images of PCA w/ GLS weighting scores on 1, 2 and 3. Similar to MAF results. Objective function ~similar, but PCA scores and loadings orthogonal.



contrasted image



259

EIGENVECTOR RESEARCH INCORPORATED

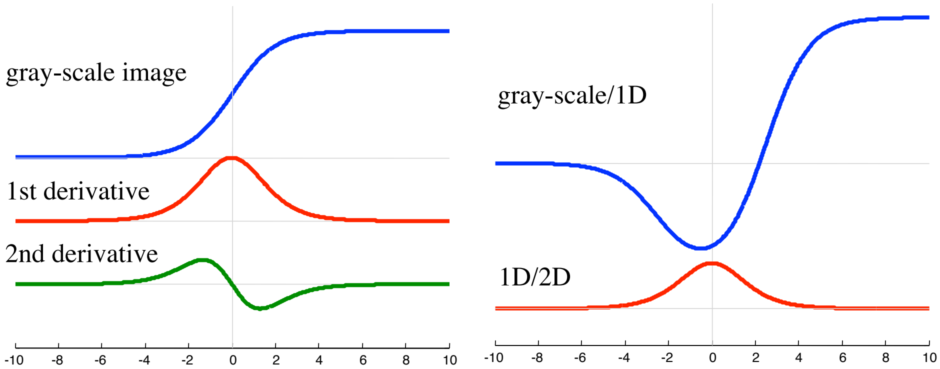
Maximum Difference Factors (MDF)

- In MDF the signal covariance corresponds to the first derivative across the spatial dimensions.
  - in MAF the first difference is the clutter
- The clutter corresponds to the second derivative across the spatial dimensions.
- Gives a multivariate analysis estimate of edges in an image.
  - analogous method available for GLS weighting w/ PCA

260

EIGENVECTOR RESEARCH INCORPORATED

MDF

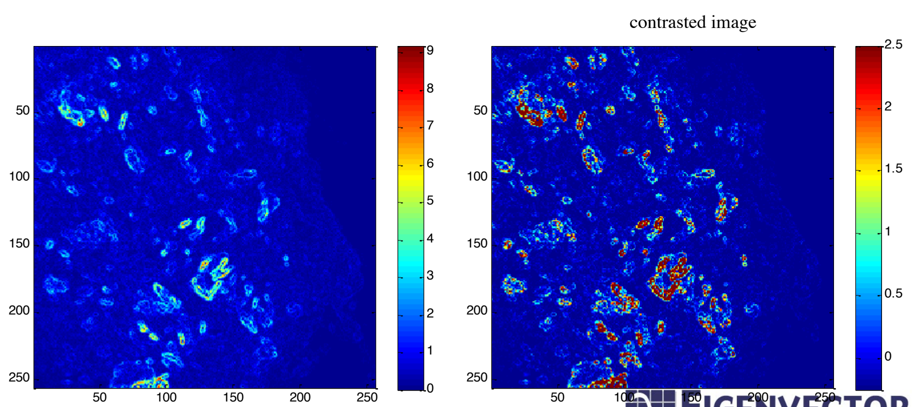


MDF finds locations in the image where the ratio of first to second derivative is a maximum

EIGENVECTOR RESEARCH INCORPORATED

MDF Results

Scaled image of MDF scores on 1
sqrt(dx^2 + dy^2)

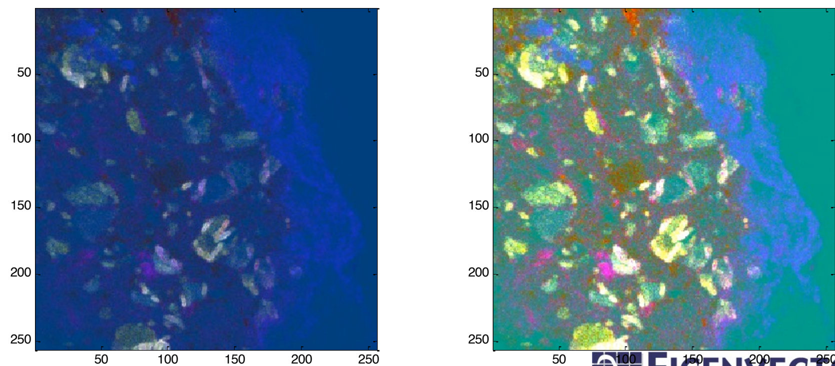


EIGENVECTOR RESEARCH INCORPORATED

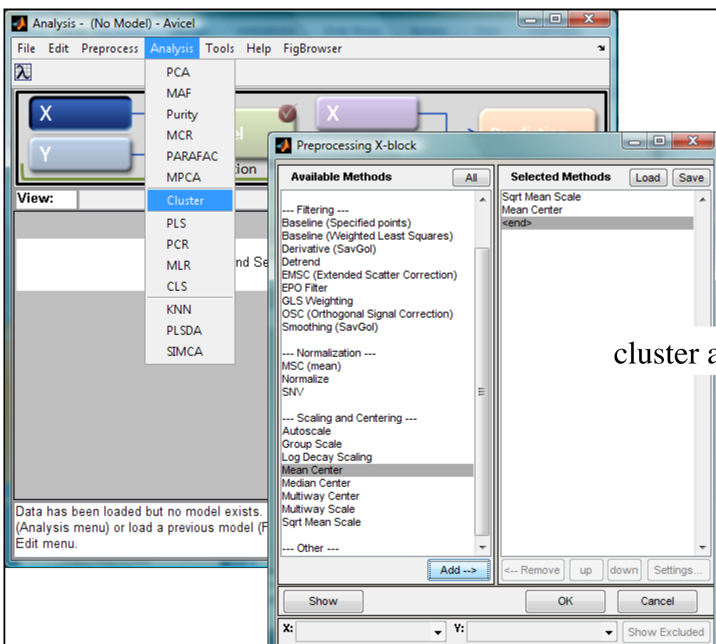
MAF+MDF Results

RGB image of MAF scores 1, 2 and 3 + MDF scores on 1
sqrt(dx^2 + dy^2)

contrasted image

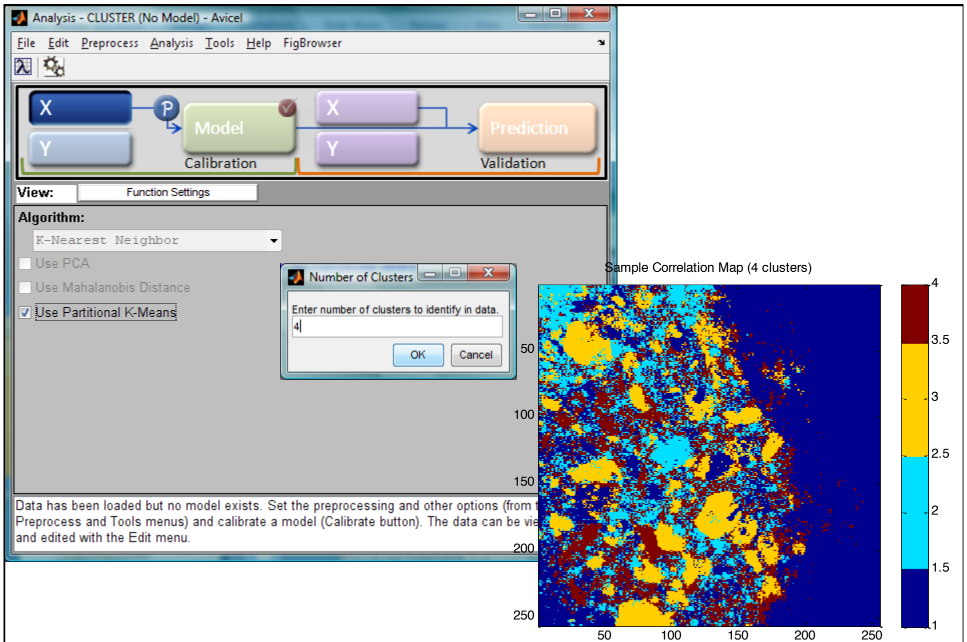


EIGENVECTOR RESEARCH INCORPORATED



cluster analysis

EIGENVECTOR RESEARCH INCORPORATED



EIGENVECTOR RESEARCH INCORPORATED

265

MIA …

- Much more to MIA
  - linked scores plots and density plots
    - interactive exploration of the image(s)
  - image SIMCA and PLS-DA
    - classification
  - curve resolution
    - chemical identification and mapping
  - image statistical process control (ISPC) for multivariate statistical process control (MSPC)
  - …

EIGENVECTOR RESEARCH INCORPORATED

266

Outline

- Introduction
- Advanced Preprocessing
- Multivariate image analysis
- Multi-way Analysis
- Summary

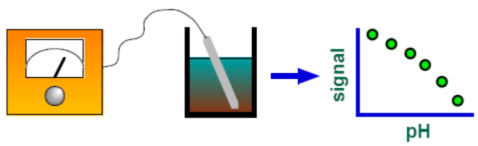
Definition of Order

- The order of a device is equal to the dimension (number of modes) of the data it produces for each sample:
- a single datum per sample → zero order
- a vector (first order tensor) per sample → first order
- a matrix (second order tensor) sample → second order
- Multi-way analysis is concerned with data with three or more modes

268

Zero Order Instrument

- The most basic instruments are zero order devices
  - produce a single datum per sample
    - pH, temperature, absorbance at a single channel
  - no way to detect errors or interferences

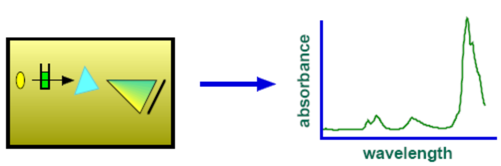


269

EIGENVECTOR RESEARCH INCORPORATED

First Order Instrument

- Many analytical instruments are first order
  - produce a vector for each sample
    - spectroscopy, LC, GC, sensor arrays
  - the presence of interferents can be detected but not corrected

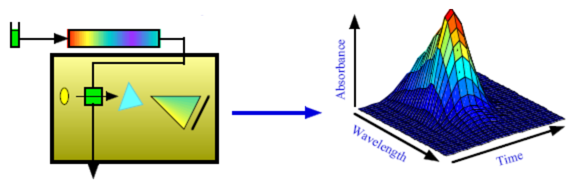


270

EIGENVECTOR RESEARCH INCORPORATED

Second Order Instrument

- Many analytical instruments are second order
  - produce a matrix for each sample
    - separation followed by spectroscopy, GC-MS, LC-UV
  - interferents can be detected and accounted for

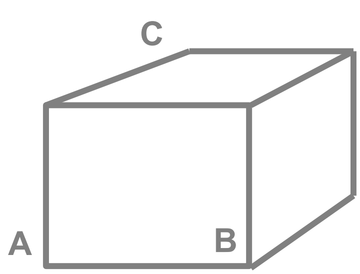


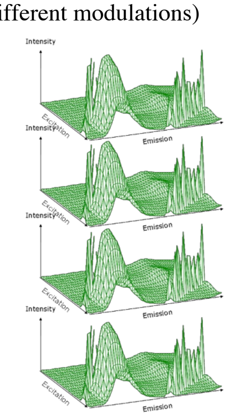
271

EIGENVECTOR RESEARCH INCORPORATED

Three-way data

- A set of ‘equivalent’ two-way matrices obtained at different occasions
- Data measured as a function of three 'things' (three different modulations)
  - E.g. samples, variables, times
- x_{ij} is a matrix element and x_{ijk} is a three-way element





272

EIGENVECTOR RESEARCH INCORPORATED

Examples

- Sensory analysis
  - Score as a function of (Food sample, Judge, Attribute)
- Process analysis
  - Measurement as a function of (Batch, Variable, time)
  - Measurement as a function of (Variable, Lag, Location)
- Image analysis
  - Pixelvalue as a function of (Sample, Image pixel, Variable)
- Experimental design
  - Response as a function of (factor 1, factor2, factor3,..)
- Spectroscopy
  - Intensity as a function of (Wavelength, Retention, Sample, Time, Location , Treatment)
- Environmental analysis
  - Measurement as a function of (Location, Time, Variable)
- Chromatography
  - Measurement as a function of (Sample, Retention time, Variable)

273

Multi-way Algorithms

- Multi-way PCA (weakly multi-way)
- Generalized Rank Annihilation (GRAM)
- Tri-Linear Decomposition (GRAM)
- PARallel FActor Analysis (PARAFAC)
- Tucker

274

PARallel FACtor analysis

PARAFAC invented in 1970 by Harshman and independently by Carroll & Chang under the name CANDECOMP. Based on a principle of parallel proportional profiles suggested in 1944 by Cattell

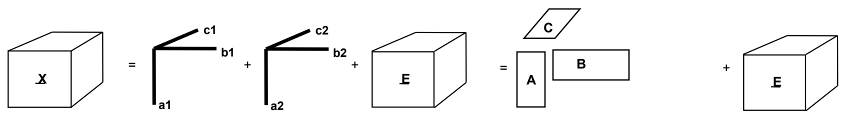
•R. A. Harshman. *UCLA working papers in phonetics* 16:1-84, 1970.
•J. D. Carroll and J. Chang. *Psychometrika* 35:283-319, 1970.
•R. B. Cattell. *Psychometrika* 9:267-283, 1944.

- PCA - bilinear model,

$$x_{ij} = \sum_{f=1}^{F} a_{if} b_{jf} + e_{ij}$$

- PARAFAC - trilinear model,

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk}$$



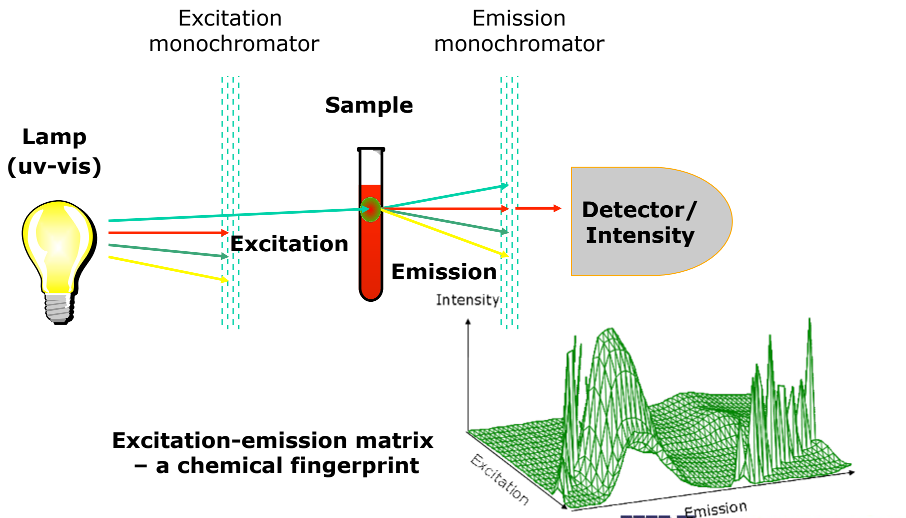
EIGENVECTOR RESEARCH INCORPORATED

Example: Excitation-Emission Fluorescence

- Use EEM measurements for quantification
  - Measure pure response for target: TRP
  - Measure response of a mixture that includes target + interferences
- Spectra are highly overlapped in both modes
- Example of second order calibration
  - the goal is to detect one analyte in the presence of unknown varying interferents using the entire EEM response
  - use PARAFAC
  - quantification in the presence of previously unseen interferents

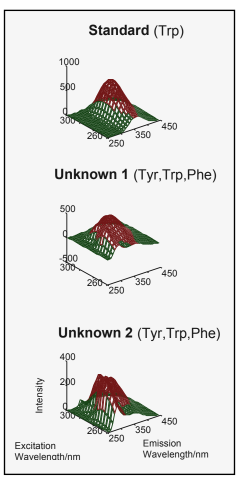
EIGENVECTOR RESEARCH INCORPORATED

Excitation Emission Fluorescence Spectroscopy



277

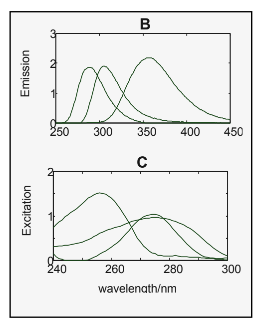
EIGENVECTOR RESEARCH INCORPORATED



Who Needs Regression?

Calibration set: *One* sample with *one* analyte (2.67 μM Trp)

Test set: Two samples with three analytes each (Trp, Tyr, Phe)



$$\begin{array}{cccc} & \textbf{Trp} & \textbf{Tyr} & \textbf{Phe} \\ \mathbf{A} = & \multicolumn{3}{c}{\begin{bmatrix} 2.67 & 0 & 0 \\ 1.52 & 1.30 & 1.29 \\ .86 & 1.12 & 1.06 \end{bmatrix}} \end{array}, \quad \textbf{reference} = \begin{bmatrix} 2.67 \\ 1.58 \\ .88 \end{bmatrix}$$

278

EIGENVECTOR RESEARCH INCORPORATED

Second Order Calibration

- The PARAFAC model estimated
  - amount of target in the test set in the presence of interferences not seen in the calibration set!
    - this is not possible with PLS
  - estimates of the response in both modes
    - allows potential library searchs
- This has enormous potential for environmental sensing and MSPC
  - Smilde, A., Bro, R., and Geladi, P., "Multi–way Analysis with Applications in the Chemical Sciences", John Wiley & Sons, New York, NY (2004).

279

Multi-way Analysis …

- Curve resolution
  - PARAFAC needs less futsing than two-way MCR
- MSPC, images, DECRA, …
- > 3 Modes
  - GCxGCxMS, sensor fusion, …

Smilde, A., Bro, R. and Geladi, P., "Multi–way Analysis with Applications in the Chemical Sciences", John Wiley & Sons, New York, NY (2004).

280

Summary

- Data analysis requires knowledge of
  - the system, physics, chemistry and math → ~~black box~~
- Advanced Preprocessing
  - uses knowledge of the clutter (GLS, ELS, etc.)
- Multivariate image analysis
  - spatial and spectral information
- Multi-way Analysis
  - measurements a function of multiple modulations giving ≥ 3 modes per sample → quant w/ unknown interferences

EIGENVECTOR RESEARCH INCORPORATED

Section Definitions 1/2

- Multivariate image analysis (MIA): Analysis of multivariate images (for many variables → hyperspectral image analysis).
- Multivariate image (MI): A data array of dimension three (or more) where the first two dimensions are spatial and the last dimension(s) is a function of another variable.
- Maximum/minimum noise fractions (MNF): Algorithm that maximizes capture of signal relative to a clutter covariance resulting in a generalized eigenvector problem.
- Maximum autocorrelation factors (MAF): MNF with the clutter covariance corresponding to the first difference of image pixels.
- Maximum autocorrelation factors (MAF): MNF with the clutter covariance corresponding to the first difference of image pixels.
- Maximum difference factors (MDF): MNF with the signal corresponding to the covariance of the first spatial derivative and clutter covariance corresponding to the second spatial derivative.
- Order: is the dimension of the data produced per sample.
- Dimension: number of modes.

EIGENVECTOR RESEARCH INCORPORATED

Section Definitions 2/2

- Multi-way analysis: Analysis of data with three or more modes.
- Multivariate image (MI): A data array of dimension three (or more) where the first
- PARallel FACto Analysis (PARAFAC): model for multi-way analysis of ≥ 3 mode data.

EIGENVECTOR RESEARCH INCORPORATED