Using Clutter to Improve Models

Barry M. Wise

Eigenvector Research, Inc.

Manson, WA USA

EIGENVECTOR RESEARCH INCORPORATED

Abstract

Clutter, defined as the confounding effects of interfering chemical species, physical effects, noise and instrument non-idealities, is present in all measurements. Sources of clutter include variation in chemical interferents, physical effects such as scattering due to particles, changes in temperature or pressure, instrument drift, detector non-linearity, as well as non-systematic random noise. The effect of clutter on models for sample classification or regression can be mitigated through use of a clutter model. These models can be derived in a number of ways such as combined class-centered data, background characterization or y-block gradient. Once obtained, they can be used to construct filters to be used in preprocessing, such as Generalized Least Squares Weighting, (GLSW), and External Parameter Orthogonalization (EPO). Clutter models can also be used directly with alternative model forms based on Classical Least Squares (CLS) such as Extended Least Squares (ELS). This talk discusses methods for obtaining clutter models and demonstrates their use in a number of applications.

Over the past dozen years, a number of powerful spectral analysis methods have been published which make use of orthogonalization (*i.e.* projection followed by weighted subtraction) of interferences or "clutter." These filtering methods provide a means to mitigate the effect of interferences arising from background chemical or physical species, instrumental artifacts, systematic sampling errors and instrument or system drift. They have been used very effectively with complex biological systems, remote sensing applications, chemical process monitoring and calibration transfer problems.

This class of methods includes Orthogonal Partial Least Squares (O-PLS), External Parameter Orthogonalization (EPO), Dynamic Orthogonal Projection (DOP), Orthogonal Signal Correction (OSC), Constrained Principal Spectral Analysis (CPSA), Generalized Least Squares Weighting (GLSW), and Science Based Calibration (SBC) among others. All are based on the orthogonalization premise and each touts a unique ability to improve model performance, robustness, and/or interpretability.

EIGENVECTOR RESEARCH INCORPORATED

Outline

- What is clutter?
- Orthogonalization filters
- How to get a clutter models
- Ways to deal with clutter
- Examples

EIGENVECTOR RESEARCH INCORPORATED

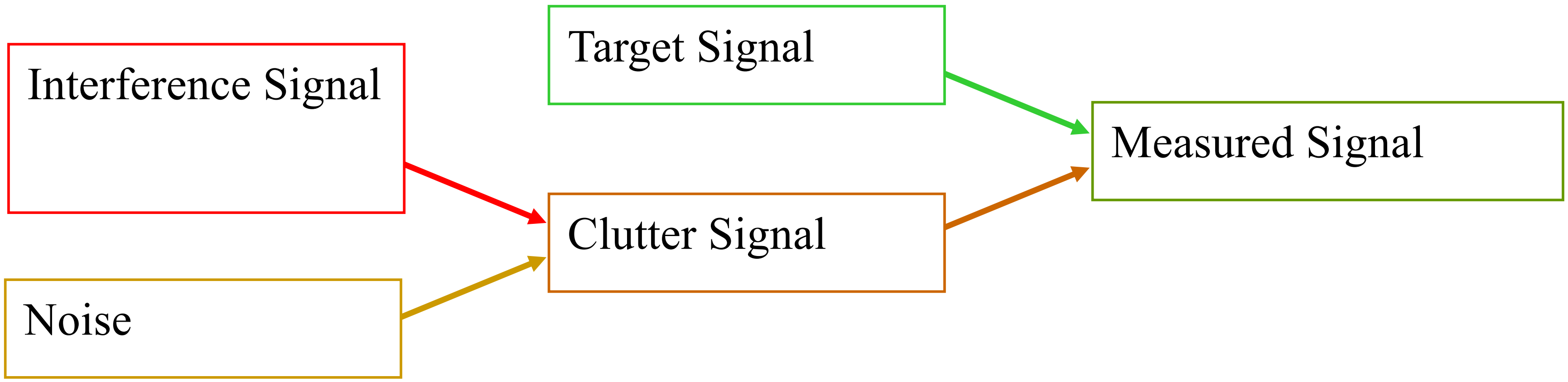
What is "Clutter?"

- A confused multitude of things: a condition in which things are not in their expected places
- Radar Clutter Definition: (DOD, NATO) Unwanted signals, echoes, or images on the face of the display tube, which interfere with observation of desired signals.
- Variations in the signal (*e.g.* spectra) not due to the factor (*e.g.* analyte) of interest due to systematic or random effects

EIGENVECTOR RESEARCH INCORPORATED

Measured Signal

- Clutter is present in all measurements
  - X-block, Y-block



Sources of Clutter

- Systematic background variability
  - in the system being sensed
    - Interfering analytes not of interest
    - Changes in particle size distribution
    - T, P changes,
    - Variable sample matrix, *e.g.* pH
  - due to physics of instrument
    - Drift, optics clouding
    - Instrument maintenance
    - Variable baseline or gain
- Non-systematic random noise
    - homoscedastic, heteroscedastic

Orthogonalization Filters

- Remove clutter from data which interfere with signal of interest
- Filters return spectra with clutter “removed”
- “Hard” orthogonalization is projection of a subspace out of the data
- "Soft” orthogonalization is deweighting but not outright complete subtraction

EIGENVECTOR RESEARCH INCORPORATED

Some Examples Using Orthogonalization Filters
(by Eigenvector)

- *In vivo* Tissue identification with NIR probe
- Cancer detection using *in vivo* fluorescence
- Identification of arthlesclerosis in artery walls using NIR
- Determination of hydroxide concentration in high-concentration aqueous ion solutions using Raman spectroscopy
- Identification of chemical species in remote sensing

EIGENVECTOR RESEARCH INCORPORATED

SOME Orthogonalization Filters

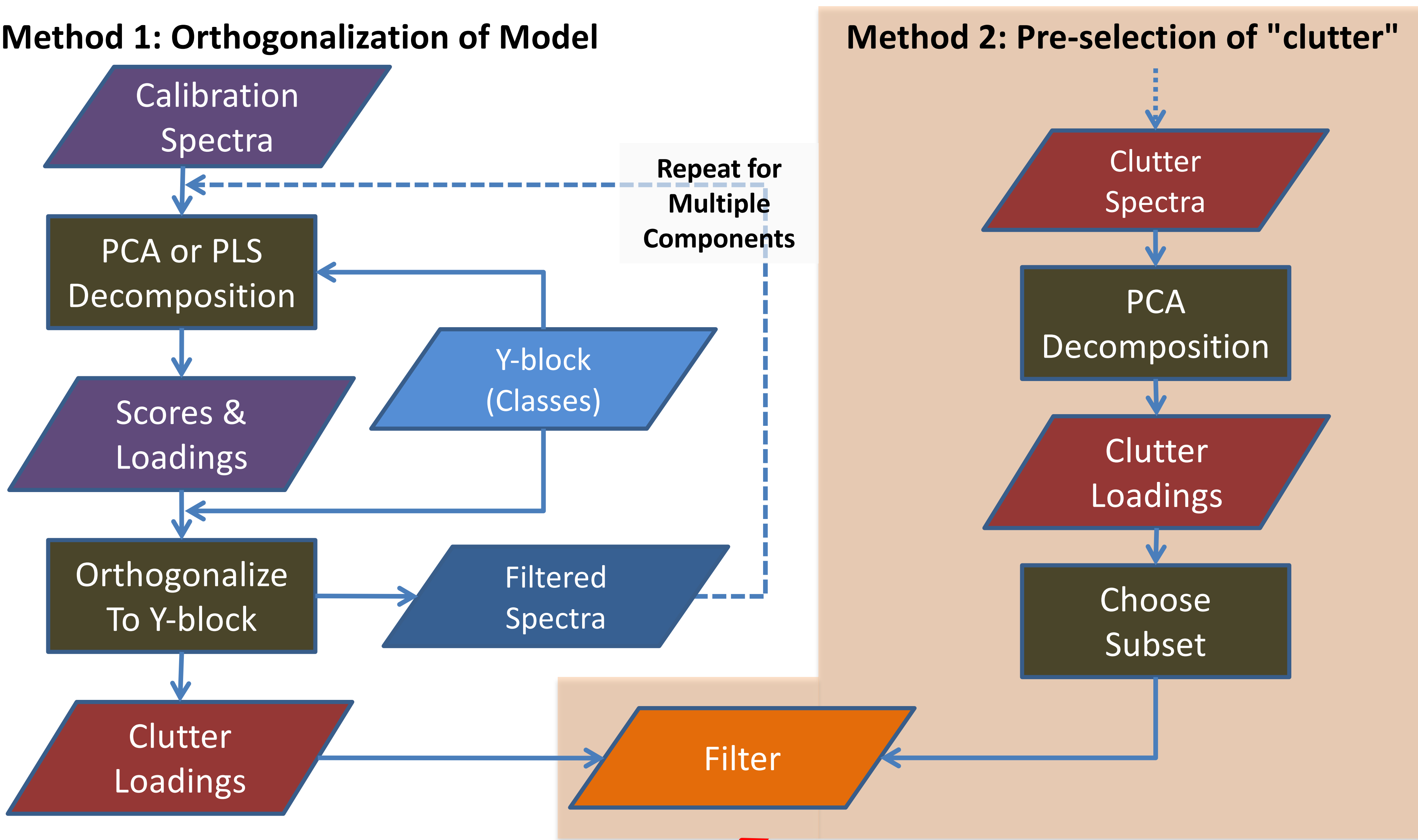
Method 1: Orthogonalization of Model

- OSC – Orthogonal Signal Correction (Wold et. al. 1998)
- OPLS – Orthogonal PLS (Trygg, Wold 2002 , patented)
- MOSC – Modified OSC (POSC - Feudale, Tan, S. Brown 2003)
- CPSA - Constrained Principal Spectral Analysis (J. Brown 1990 , patented)
- EPO – External Parameter Orthogonalization (Roger et. al 2003)
- GLS – Generalized Least Squares (Aitken 1935, Martens et. al. 2003)
- SBC – Science Based Calibration (Marbach 2005, patented)
- EMSC – Extended Multiplicative Scatter Correction (Martens, Stark)
- ELS/EMM – Extended Least Squares/Extended Mixture Model

Method 2: Pre-selection of "clutter"

EIGENVECTOR RESEARCH INCORPORATED

Two General Approaches



Focusing on this

EIGENVECTOR RESEARCH INCORPORATED

Pre-selection Methods…

- CPSA - Constrained Principal Spectral Analysis
- EPO – External Parameter Orthogonalization

  - Identical
  - Choose # of PCs

- GLS – Generalized Least Squares
- SBC – Science Based Calibration

  - Quite similar
  - Down-weight by scale of eigenvalues

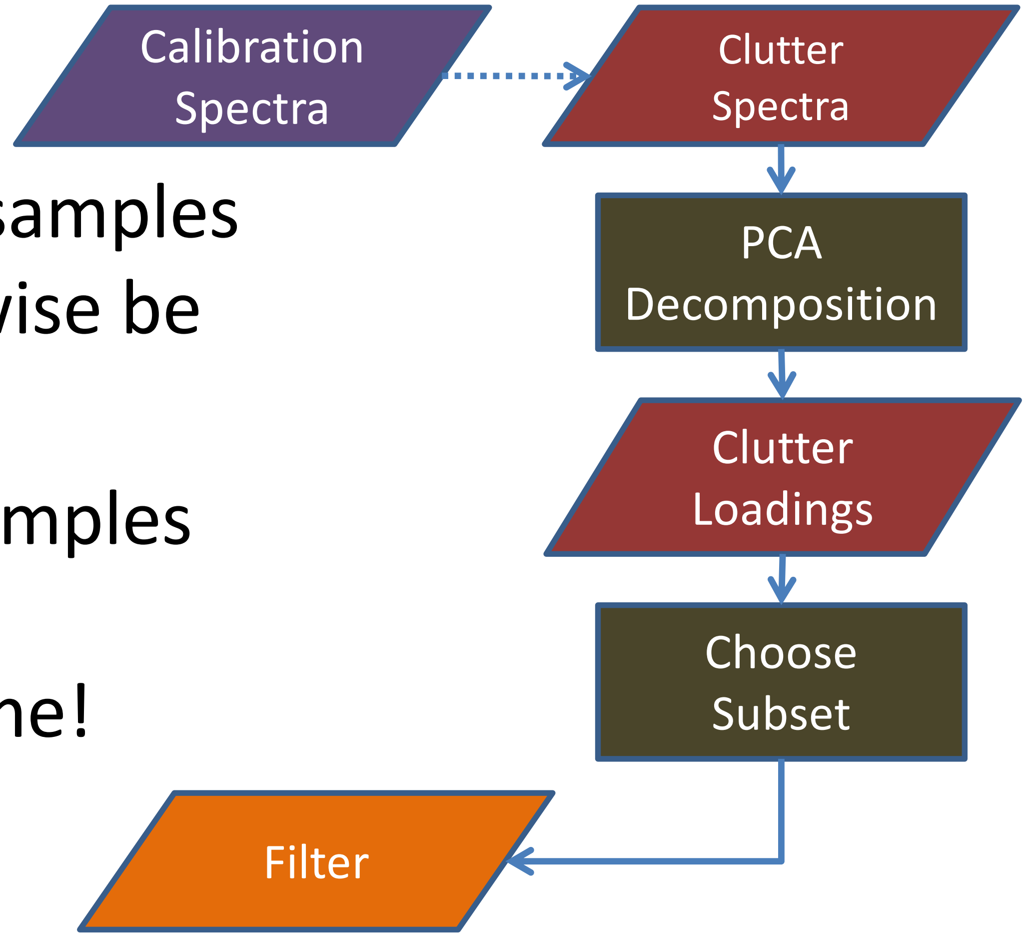
- EMSC – Extended MSC
- EMM/ELS – Extended Mixture Model

  - CLS type models

Pre-selecting Clutter

How to get clutter?

- Look at differences in samples which should otherwise be the same.
- In classification – all samples within a class should nominally be the same!
- Use Calibration itself!



More on How to Get Clutter

- Pure component spectra of known interferences
- Subspace spanned by
  - samples where analyte of interest is not present
  - variation in data that is all of the same class
  - repeat measurement of blanks
  - off-target pixels in remote sensing
- Make it up! *e.g.* polynomial baseline shapes

EIGENVECTOR RESEARCH INCORPORATED

Y-gradient Method

- Sort samples by **y** (reference) values
- Take differences between adjacent samples
- Weight **X**-differences by inverse of difference in **y** values
- Deweight by covariance of differences (GLS) or orthogonalize against some number of PCs (EPO, ELS, EMM, PA-CLS)

Clutter Covariance

Clutter source 1 Clutter source 2

$$\mathbf{X}_c = (\mathbf{X}_{1,c} - \bar{\mathbf{x}}_{1,c}) + (\mathbf{X}_{2,c} - \bar{\mathbf{x}}_{2,c}) + \ldots$$

$$\mathbf{C} = \frac{\mathbf{X}_c^{\mathrm{T}} \mathbf{X}_c}{N-1}$$

EIGENVECTOR RESEARCH INCORPORATED

Covariance to Clutter Basis

$$\mathbf{C} = \mathbf{V}\mathbf{S}^2\mathbf{V}^{\mathrm{T}}$$

For basis choose some number of factors

$$\mathbf{B} = \mathbf{V}_{1\ldots k}$$

Covariance to GLS Weighting Matrix

$$\mathbf{C} = \mathbf{V}\mathbf{S}^2\mathbf{V}^{\mathrm{T}}$$

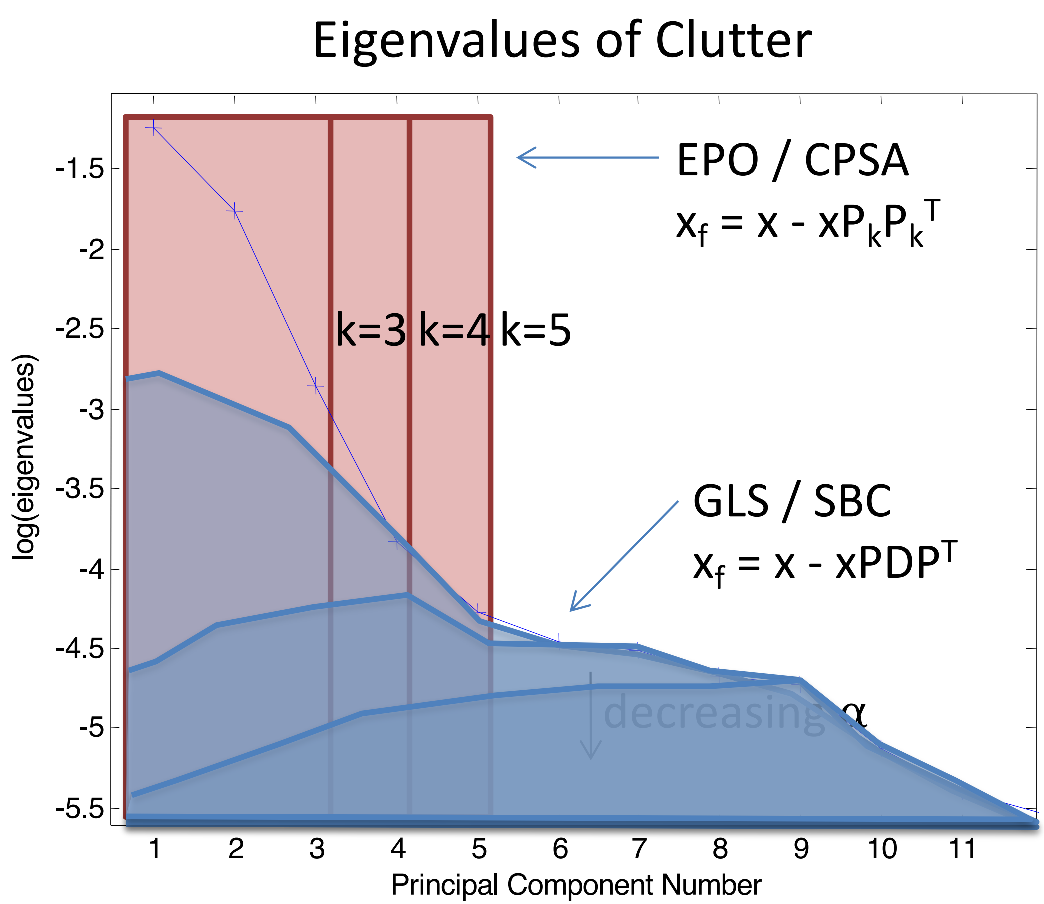
weighting matrix $$\mathbf{G} = \mathbf{V}\mathbf{D}^{-1}\mathbf{V}^{\mathrm{T}}$$

with $$d_{i,i}^{-1} = \frac{1}{\sqrt{\frac{s_{i,i}^2}{\alpha^2} + 1}}$$

Large α ➔ ∞, dimension unaffected
Small α ➔ 0, dimension eliminated

EIGENVECTOR RESEARCH INCORPORATED

Choosing Components

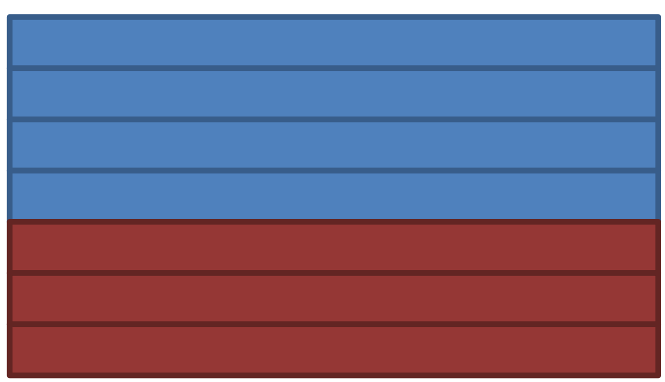


One adjustable parameter in each method

Other Similar Pre-selection Filters…

- Extended Mixture Model (Extended Least Squares) orthogonal filtering for Classical Least Squares (CLS) models!

Target (Calibration) Spectra



S_{target}

$S_{clutter}$

Clutter Spectra

$$c = xS(S^TS)^{-1}$$

Pseudo-inverse is an orthogonalization!

Equivalent to full-rank EPO / CPSA model

EIGENVECTOR RESEARCH INCORPORATED

Extended Multiplicative Scatter Correction

- EMSC attempts to correct for scatter that appears in forms other than just linear using the extended mixture model

$$\mathbf{s}_2 = \begin{bmatrix} \mathbf{s}_{ref} & \boldsymbol{\upsilon}^2 & \boldsymbol{\upsilon} & \mathbf{1} \end{bmatrix} \begin{bmatrix} c_1 \\ \mathbf{c}_P \end{bmatrix}$$

$$\mathbf{c} = \left(\mathbf{Z}^T \mathbf{Z}\right)^{-1} \mathbf{Z}^T \mathbf{s}_2$$

$$\mathbf{s}_{2,corrected} = \left(\mathbf{s}_2 - \mathbf{P}\mathbf{c}_P\right) / c_1$$

$$\mathbf{P}_{NxK} = \begin{bmatrix} \boldsymbol{\upsilon}^2 & \boldsymbol{\upsilon} & \mathbf{1} \end{bmatrix}$$

$$\mathbf{Z}_{Nx(1+K)} = \begin{bmatrix} \mathbf{s}_2 & \mathbf{P} \end{bmatrix}$$

$$\mathbf{c} = \begin{bmatrix} c_1 \\ \mathbf{c}_P \end{bmatrix}$$

EIGENVECTOR RESEARCH INCORPORATED

EMSC

- can add spectra of known target analyte $\mathbf{S}_{\mathrm{A},NxJ}$
- can add spectra or basis of clutter $\mathbf{Q}_{NxL}$.

$$\mathbf{s}_2 = \begin{bmatrix} \mathbf{s}_{ref} & \mathbf{S} & \mathbf{P} & \mathbf{Q} \end{bmatrix} \mathbf{c}$$

$$\mathbf{c} = \left(\mathbf{Z}^T \mathbf{Z} \right)^{-1} \mathbf{Z}^T \mathbf{s}_2$$

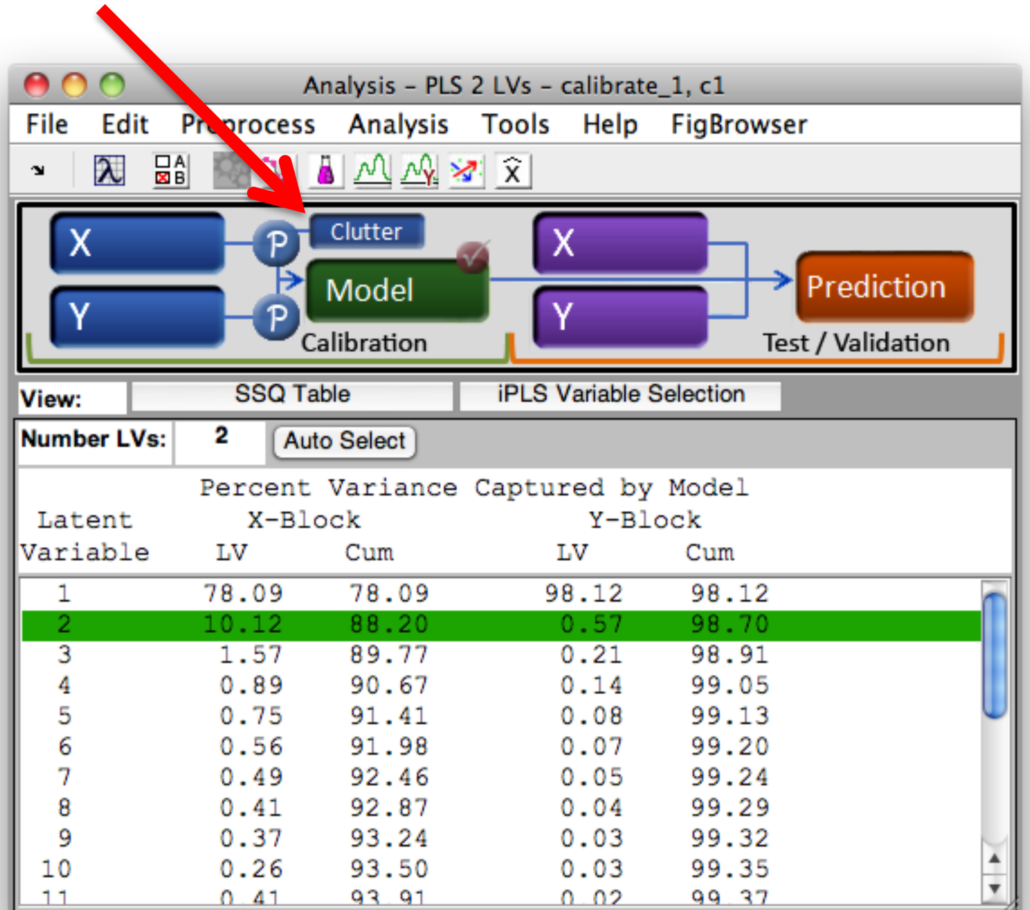
$$\mathbf{s}_{2,corrected} = \left(\mathbf{s}_2 - \mathbf{P}\mathbf{c}_P - \mathbf{Q}\mathbf{c}_Q \right) \big/ c_1$$

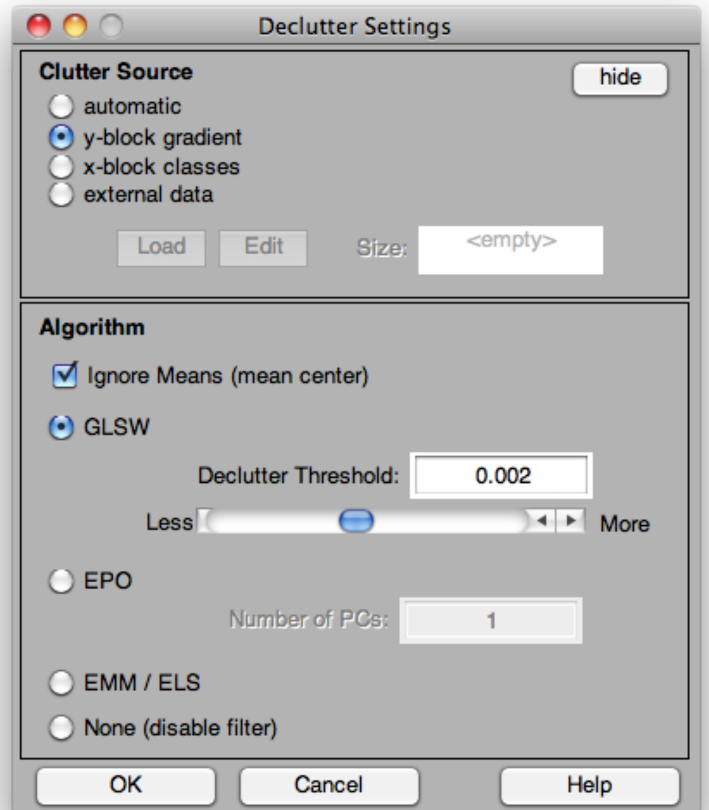
$$\mathbf{P}_{NxK} = \begin{bmatrix} \cdots & \boldsymbol{\upsilon}^2 & \boldsymbol{\upsilon} & \mathbf{1} \end{bmatrix}$$

$$\mathbf{Z}_{Nx(1+J+K+L)} = \begin{bmatrix} \mathbf{s}_{ref} & \mathbf{S}_A & \mathbf{P} & \mathbf{Q} \end{bmatrix}$$

$$\mathbf{c}^T = \begin{bmatrix} c_1 & \mathbf{c}_S^T & \mathbf{c}_P^T & \mathbf{c}_Q^T \end{bmatrix}_{1x(1+J+K+L)}$$

We think it is useful to use Clutter!

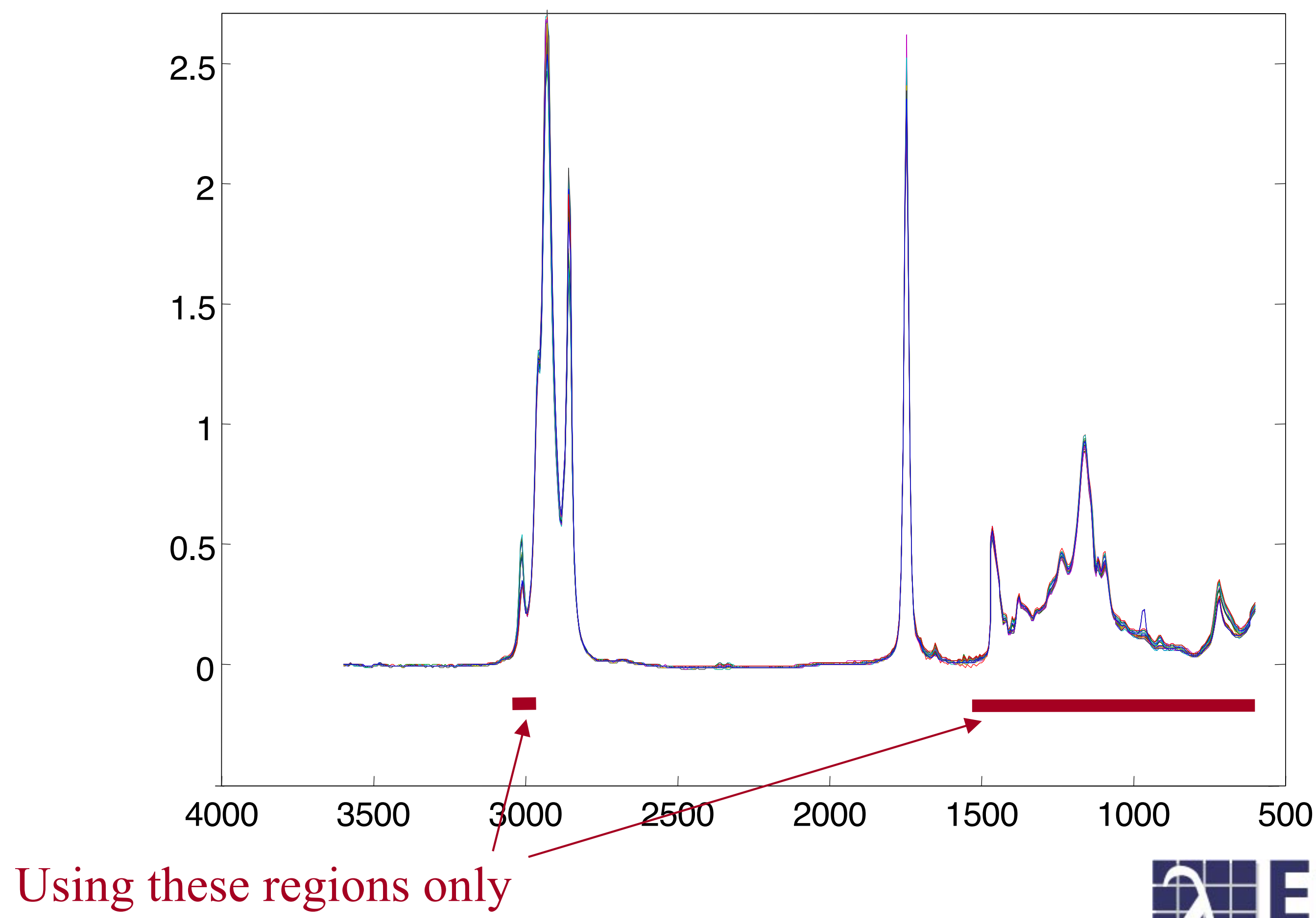




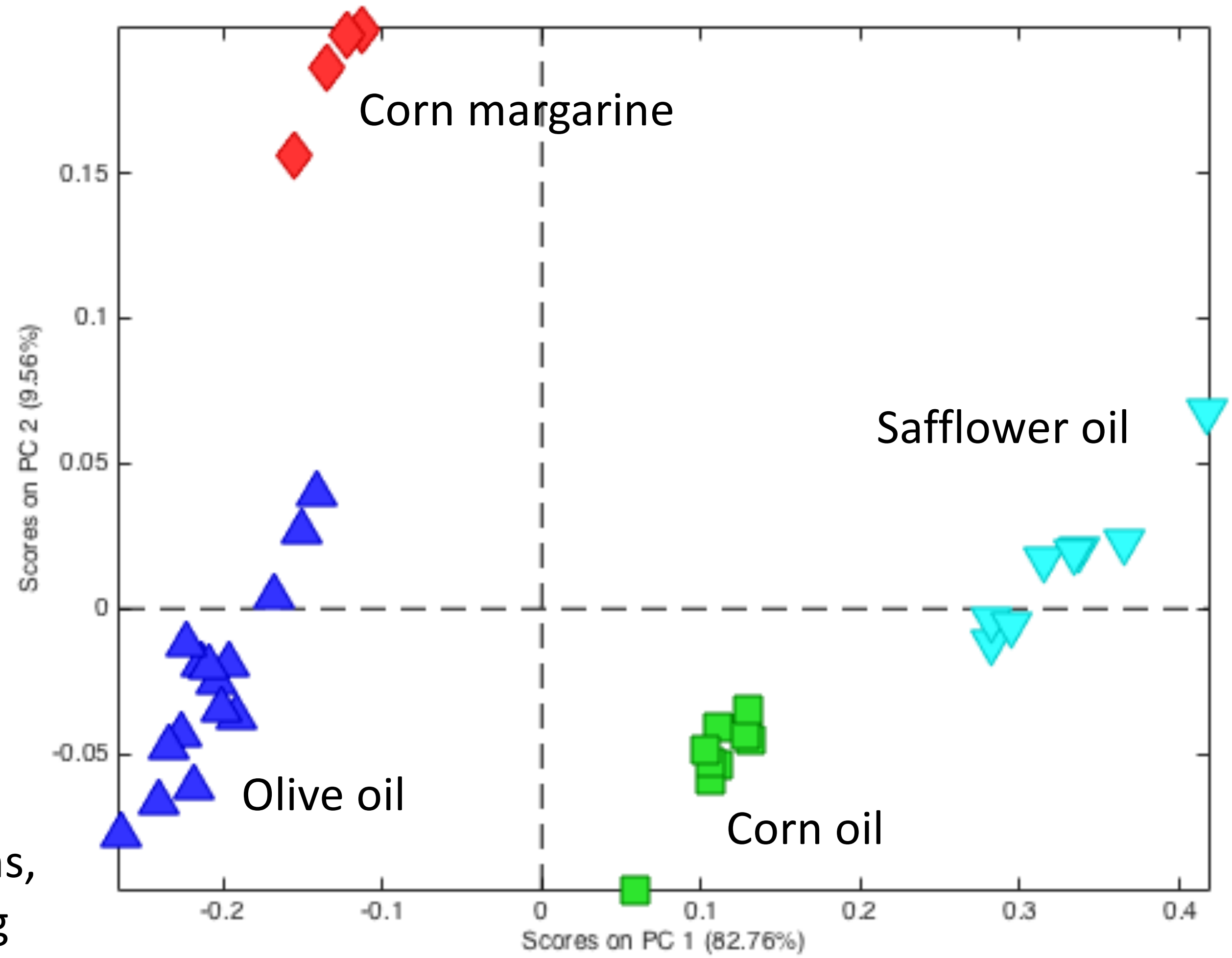
EIGENVECTOR RESEARCH INCORPORATED

Example Classification Data

- Mid-IR spectra of food grade oils
- Classify oils, detect adulterated olive oil

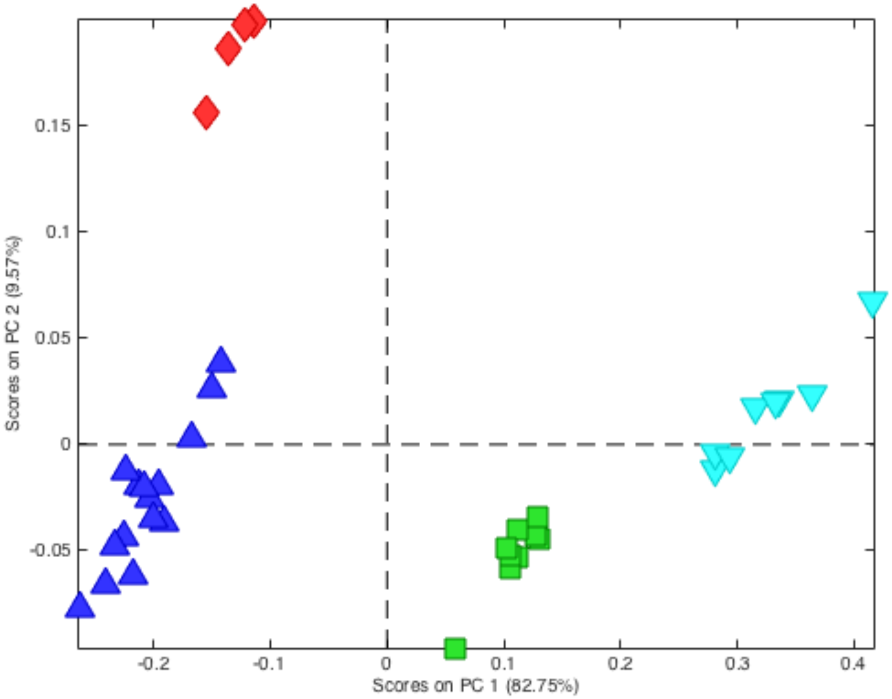


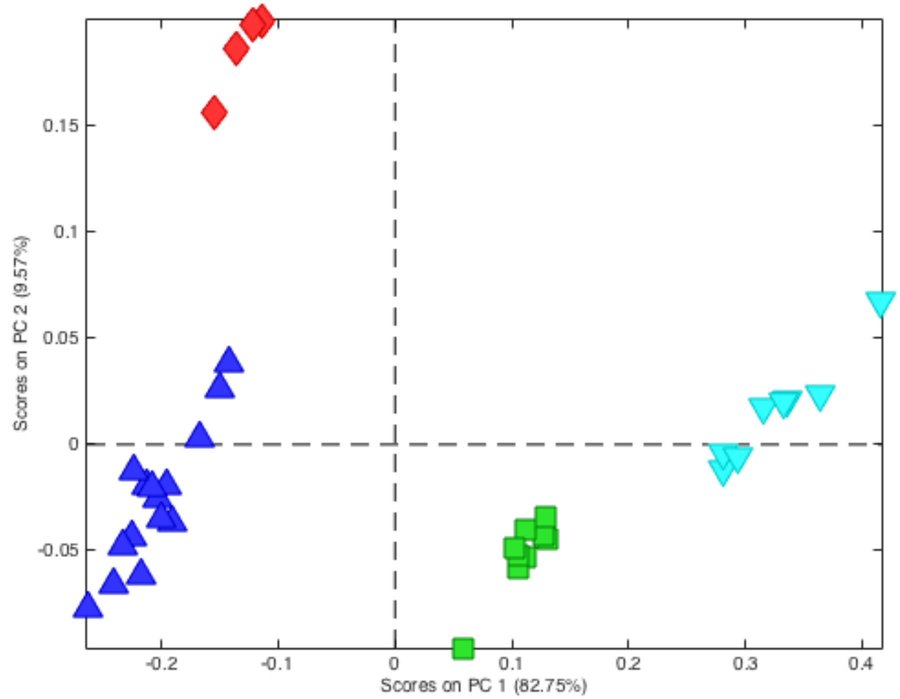
PCA Scores Plot of Oils



Selected regions, mean centering only

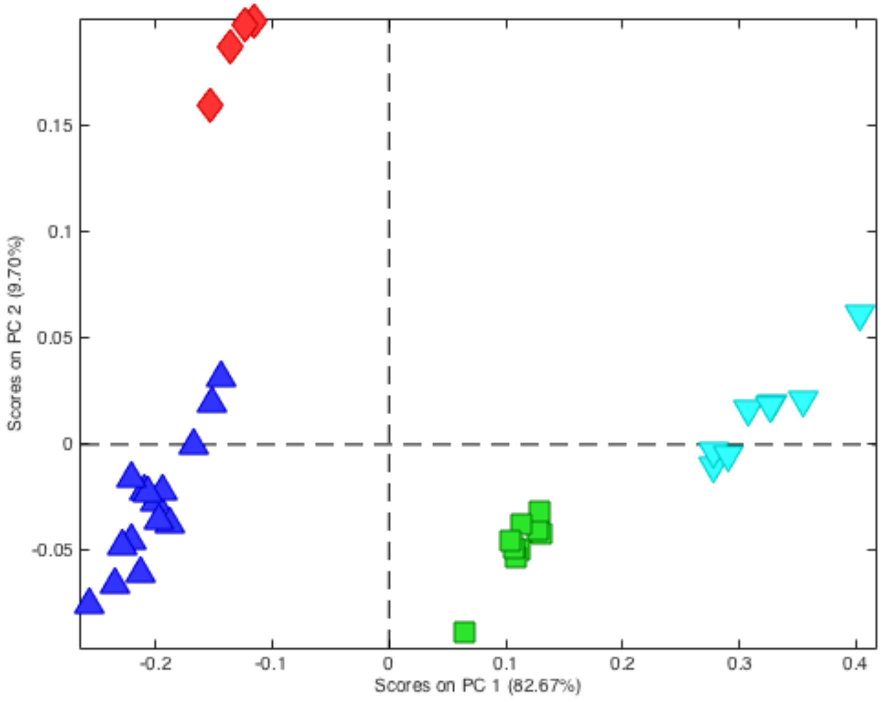
GLS α = 1

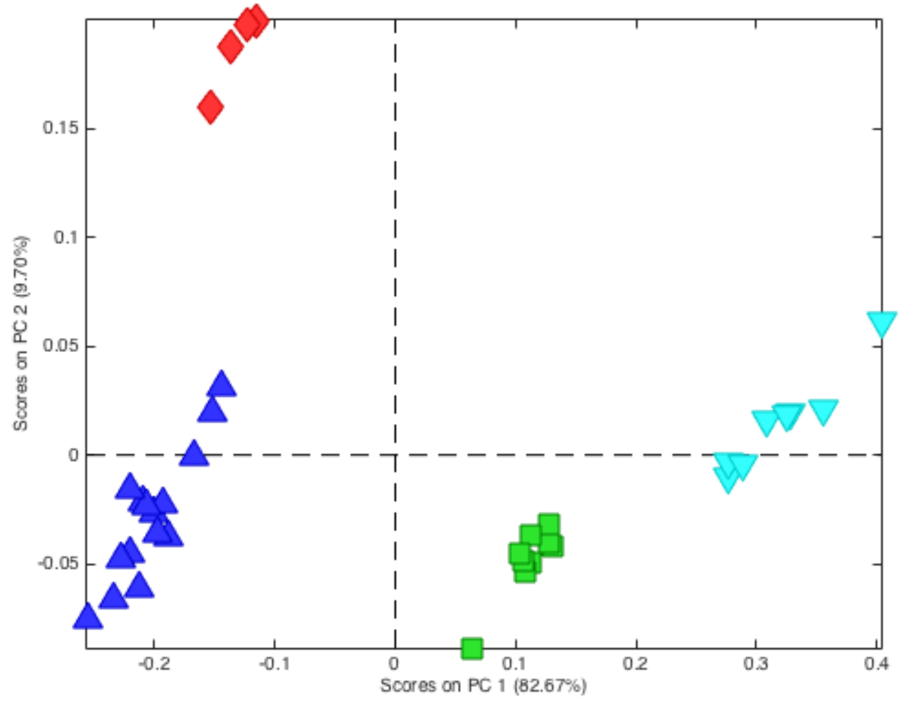




EIGENVECTOR RESEARCH INCORPORATED

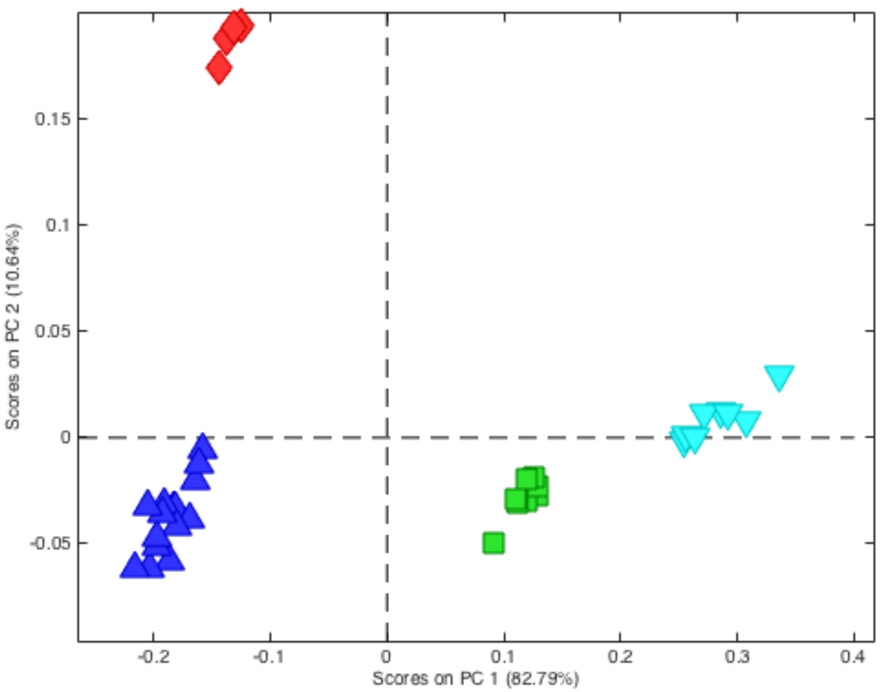
GLS $\alpha = 0.3$

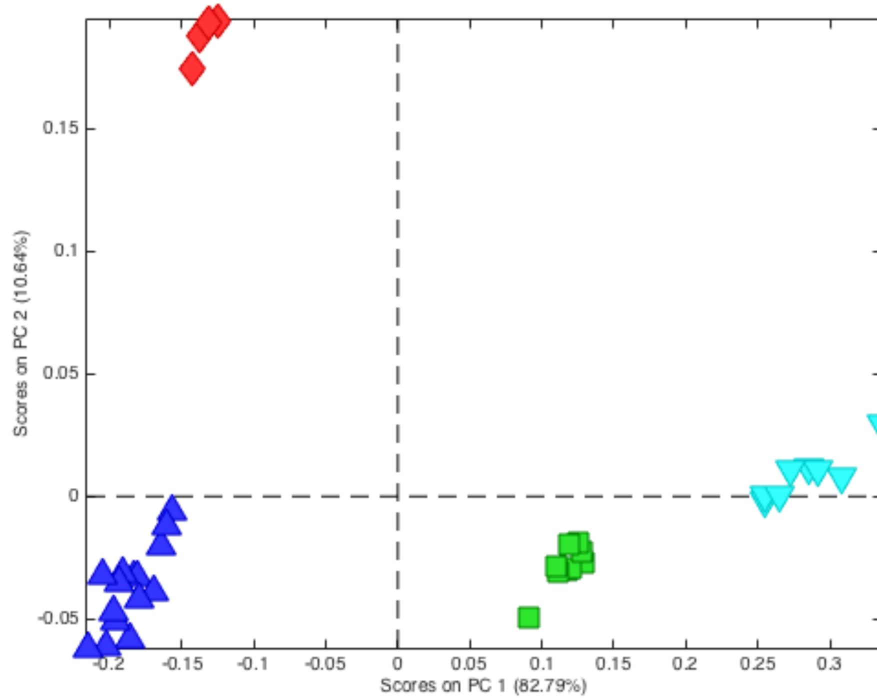




EIGENVECTOR
RESEARCH INCORPORATED

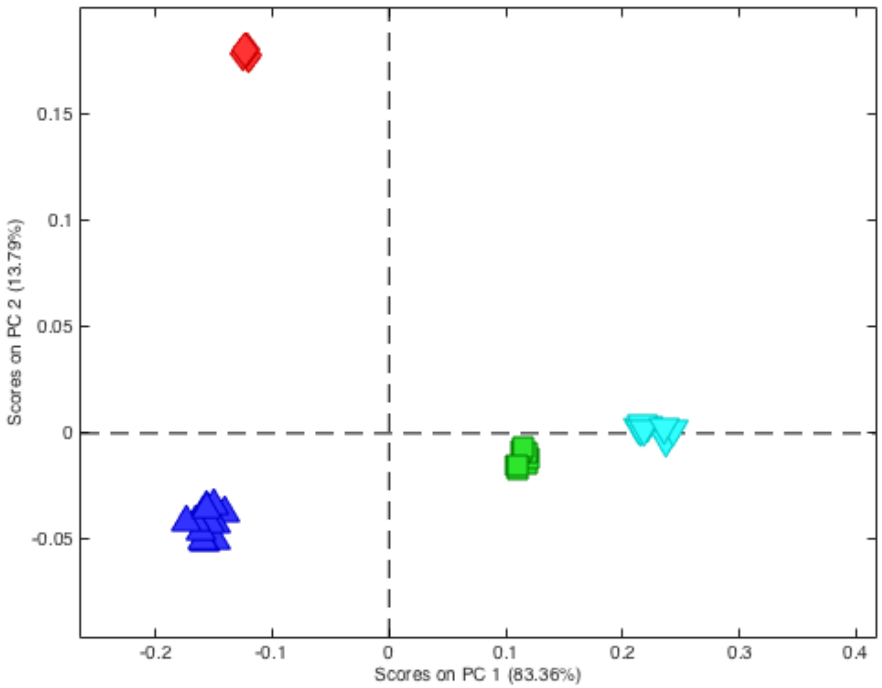
GLS $\alpha = 0.1$

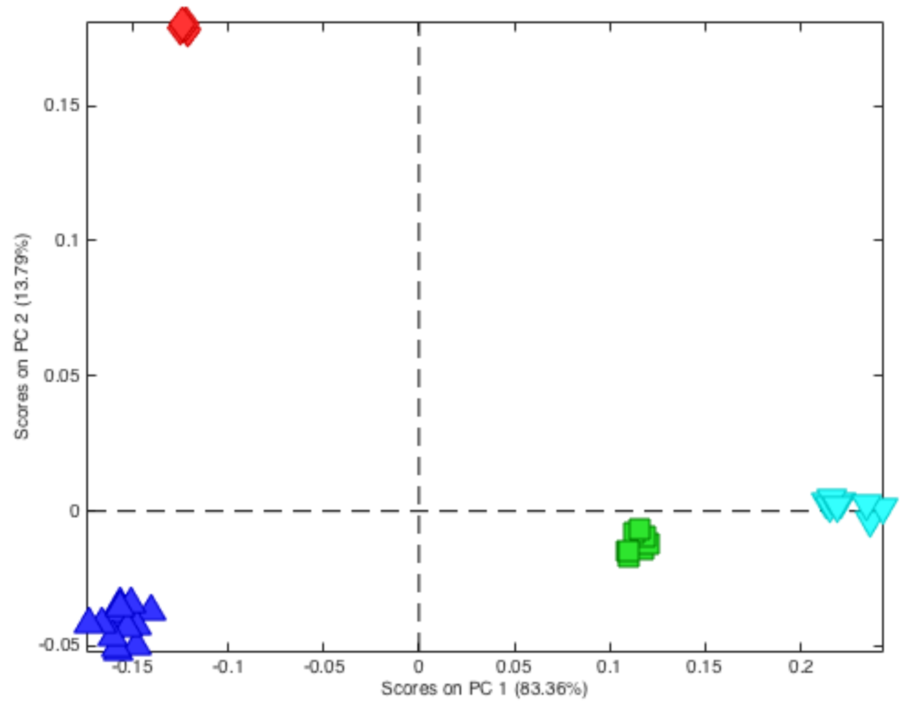




EIGENVECTOR RESEARCH INCORPORATED

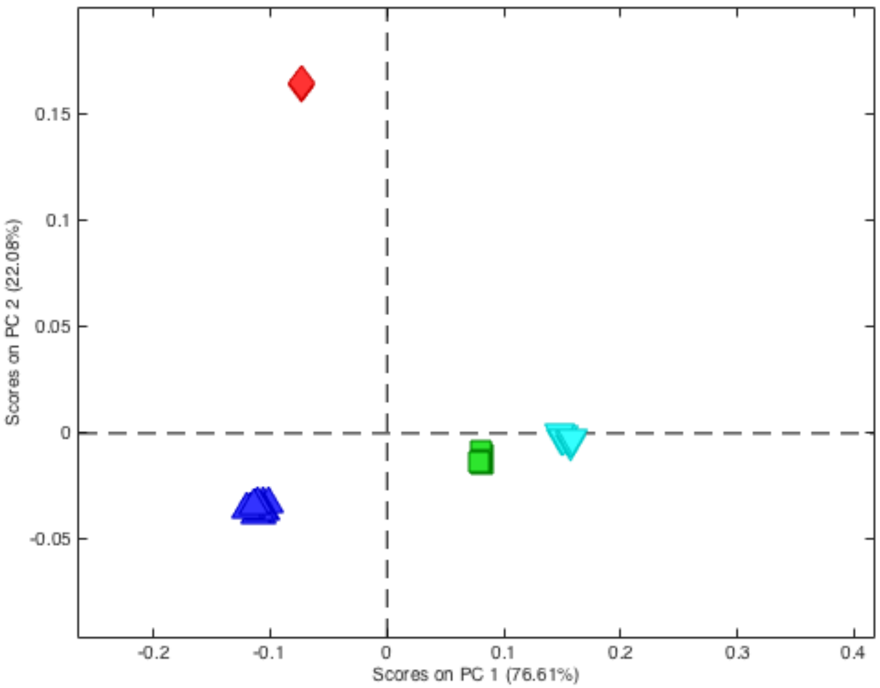
GLS α = 0.03

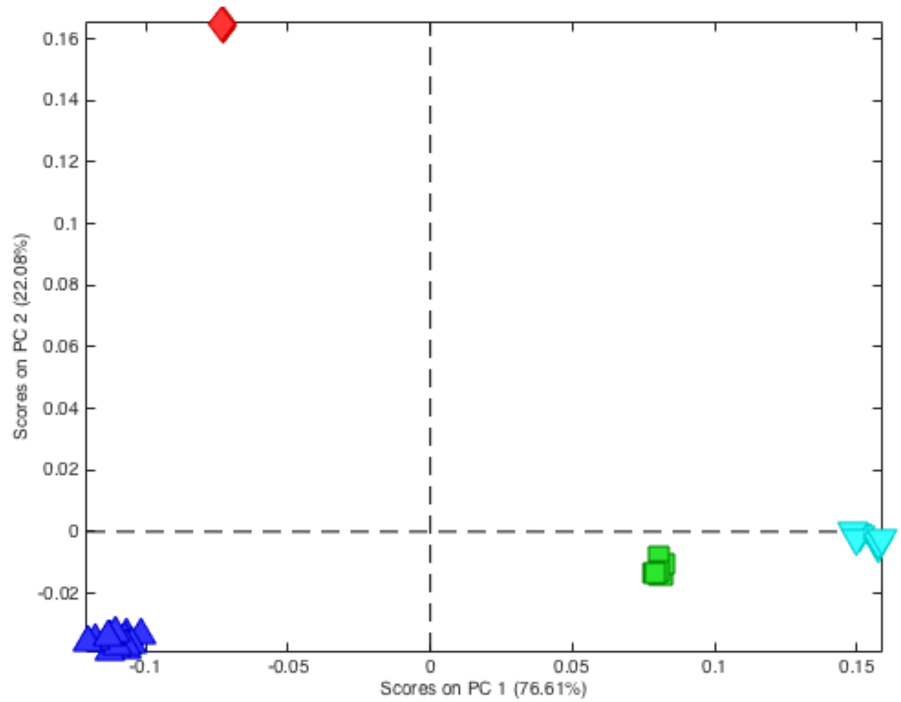




EIGENVECTOR RESEARCH INCORPORATED

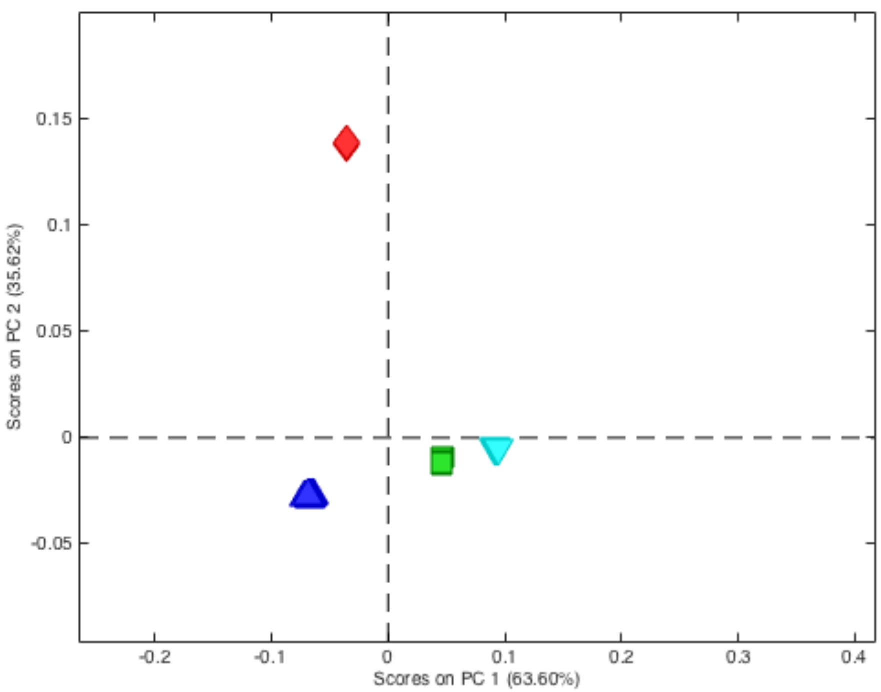
GLS α = 0.01

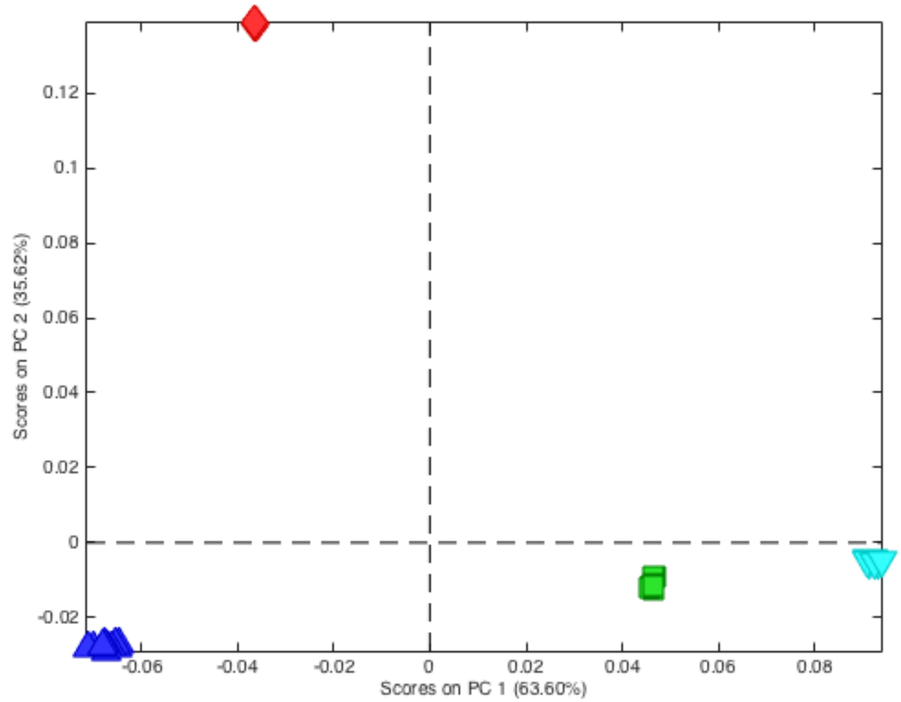




EIGENVECTOR RESEARCH INCORPORATED

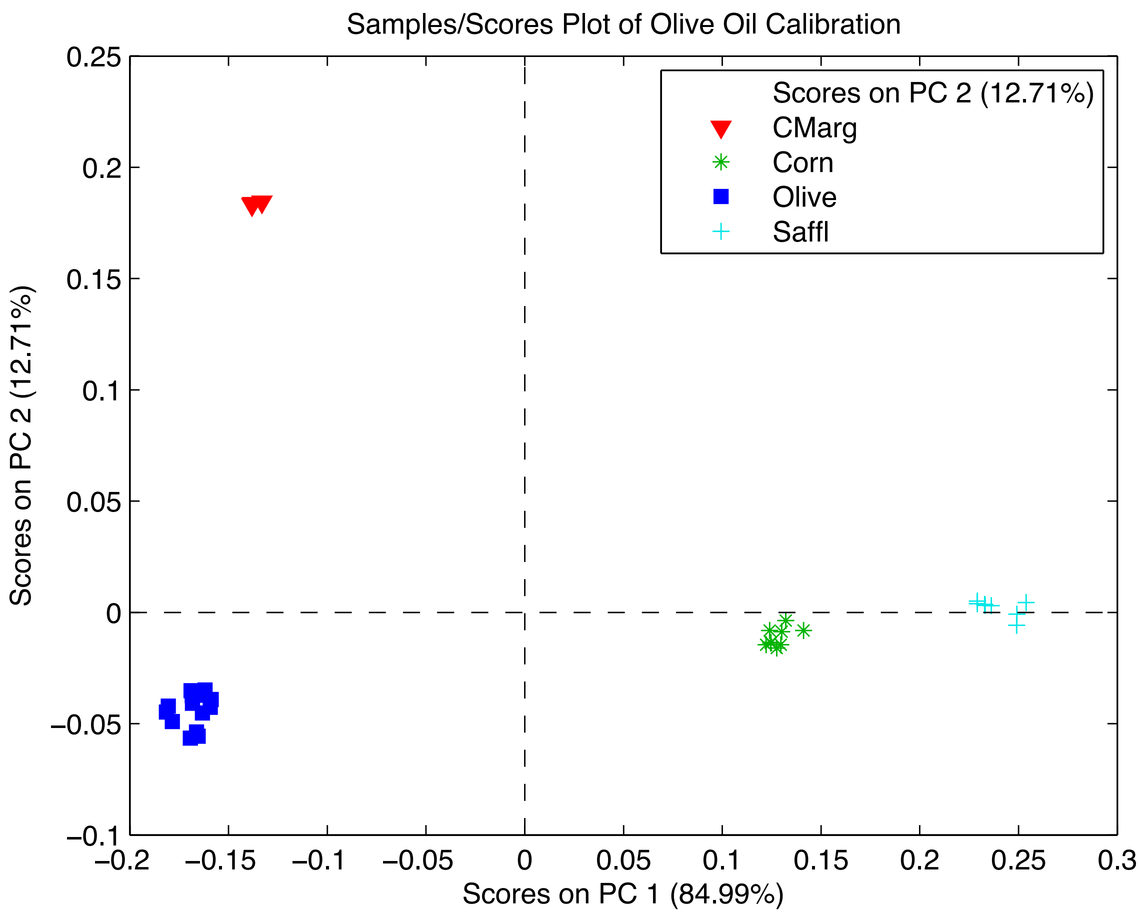
GLS α = 0.003



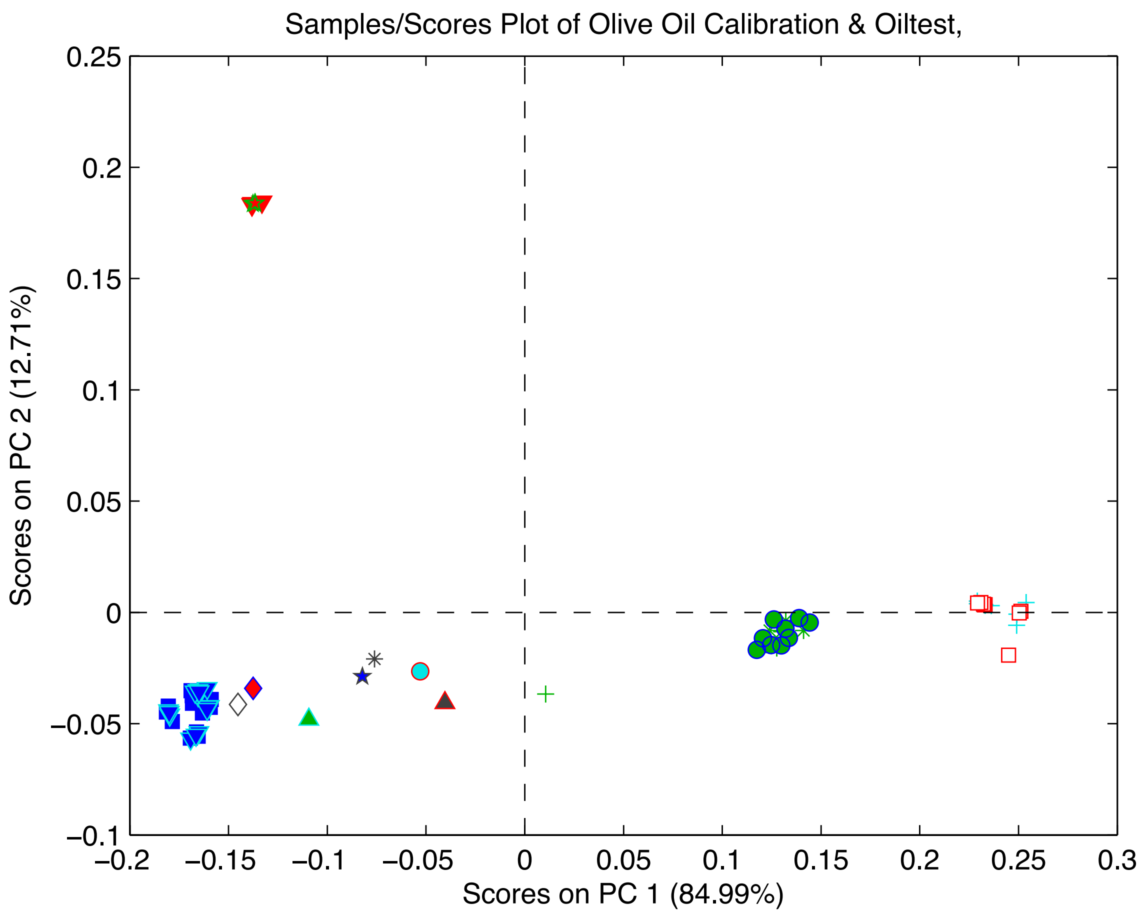


EIGENVECTOR RESEARCH INCORPORATED

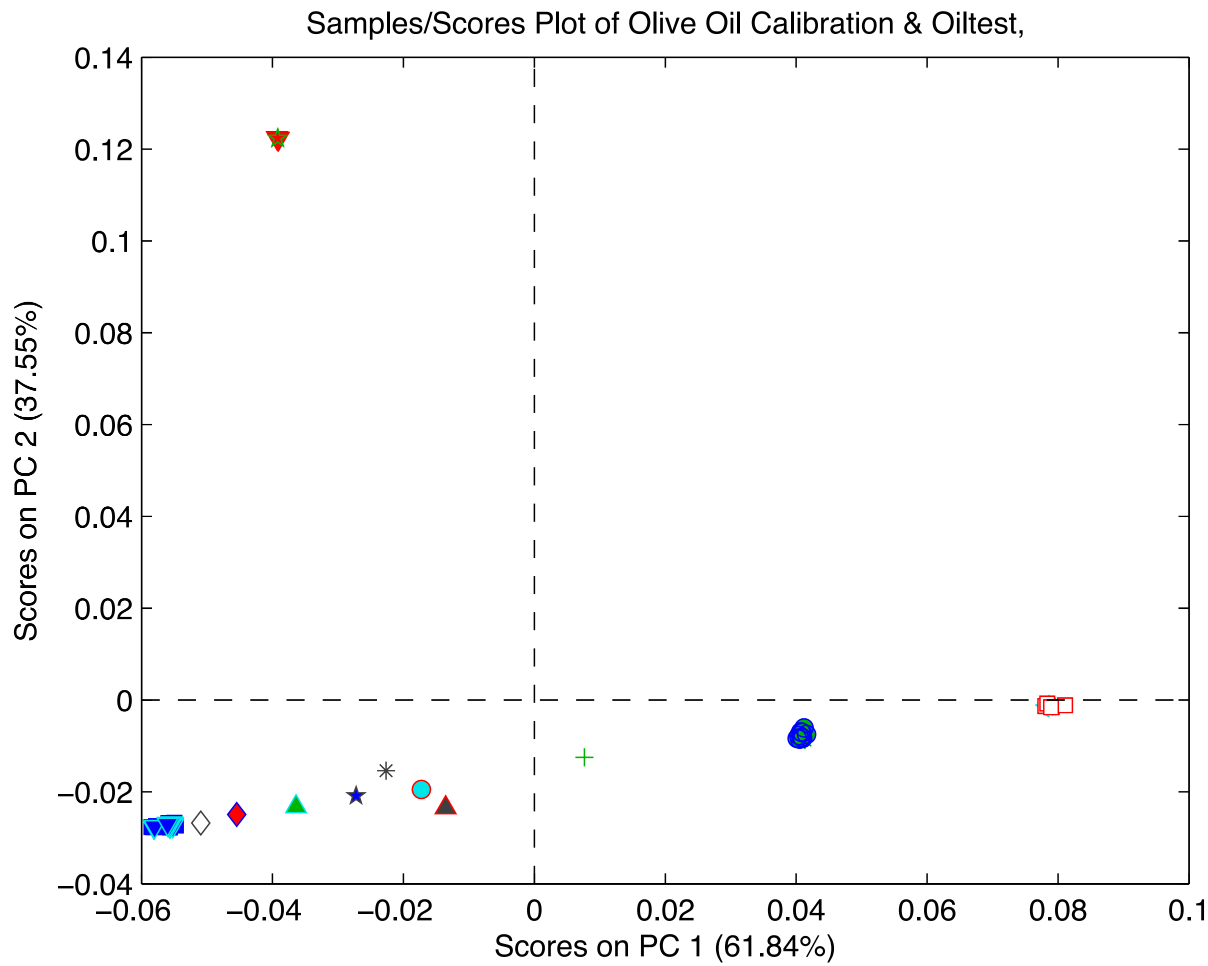
Calibration with MSC



Cal and Test with MSC

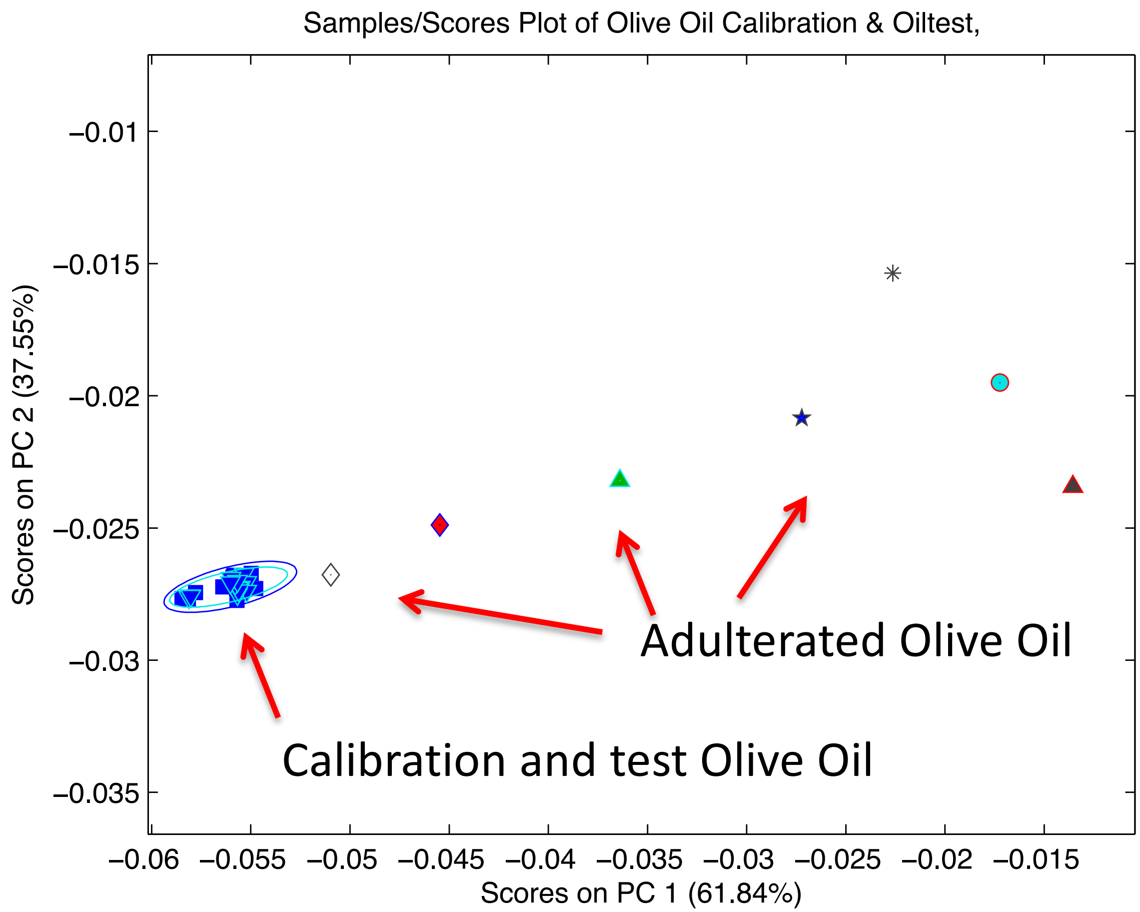


With MSC and GLS

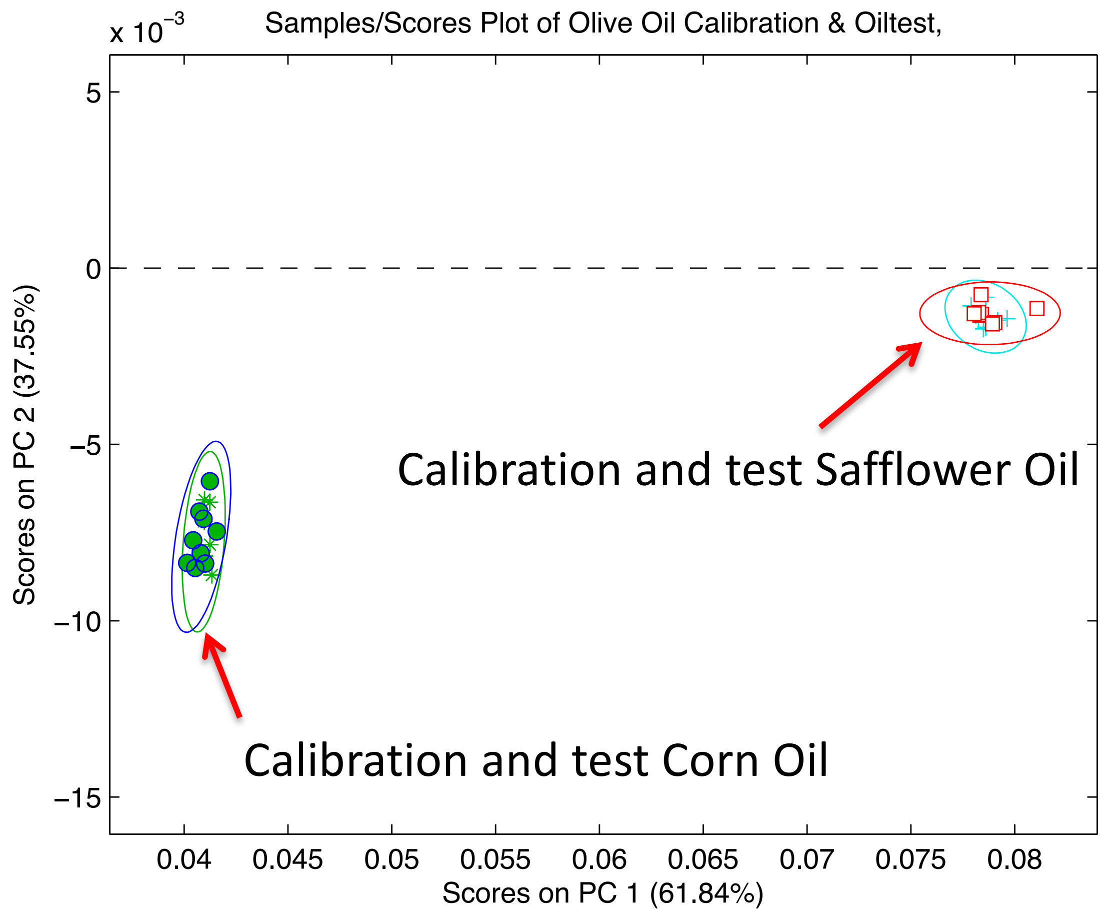


EIGENVECTOR RESEARCH INCORPORATED

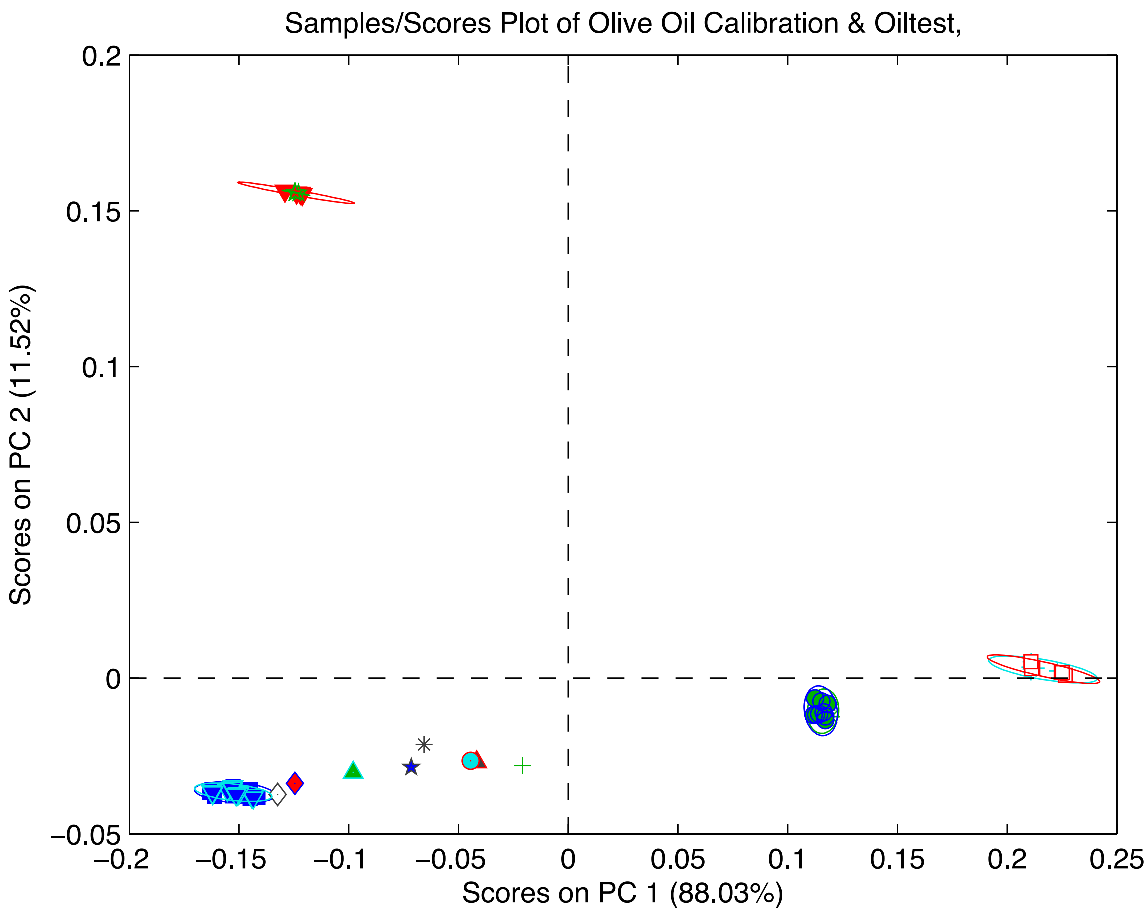
Zoom on Olive Oil



Zoom on Corn and Safflower Oil



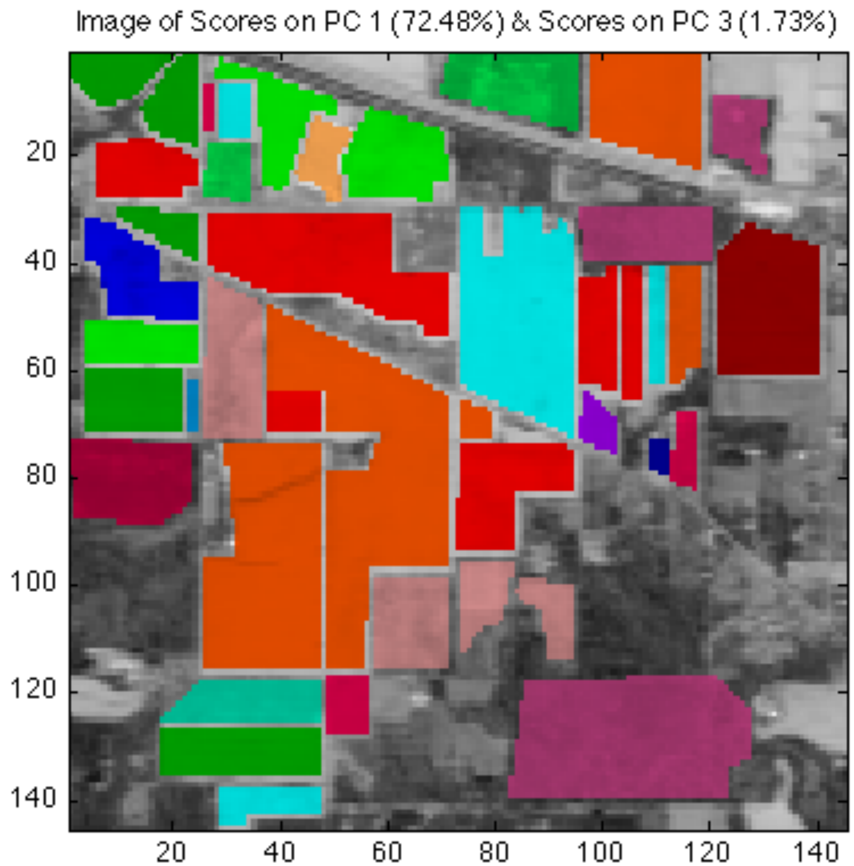
With MSC and EPO

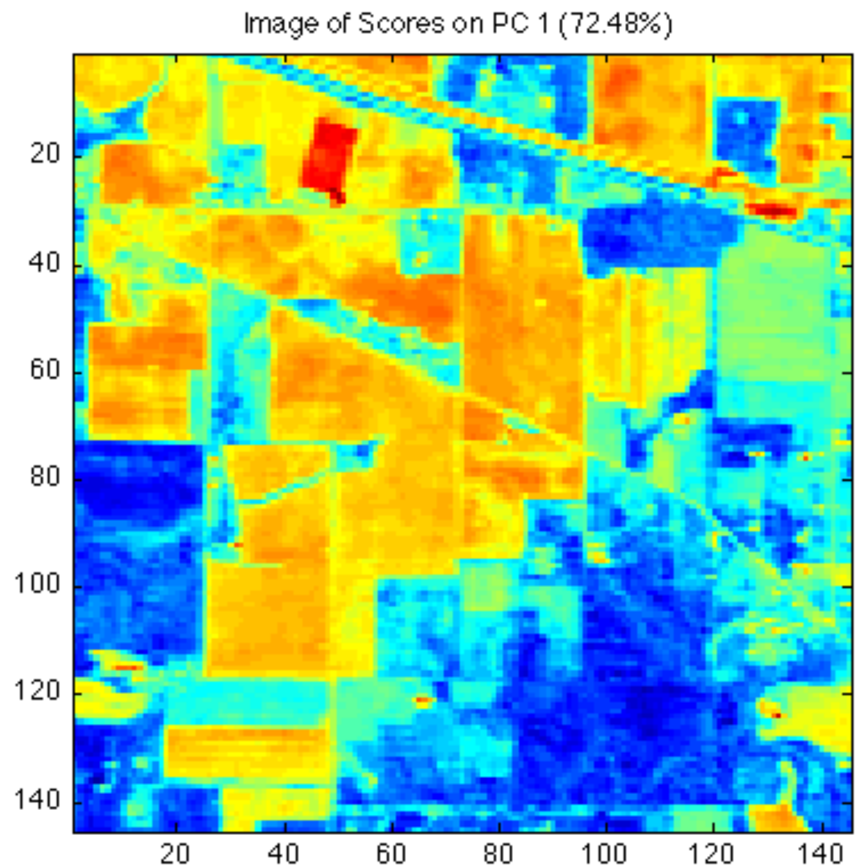


Indian Pines Data

- Classic image data set used in many publications
- Crop area near West Lafayette, Indiana
- Ground truth identified 16 know crop areas
- Data from AVIRIS: Airborne Visible/Infrared Imaging Spectrometer
- 220 channels, 400-2500nm

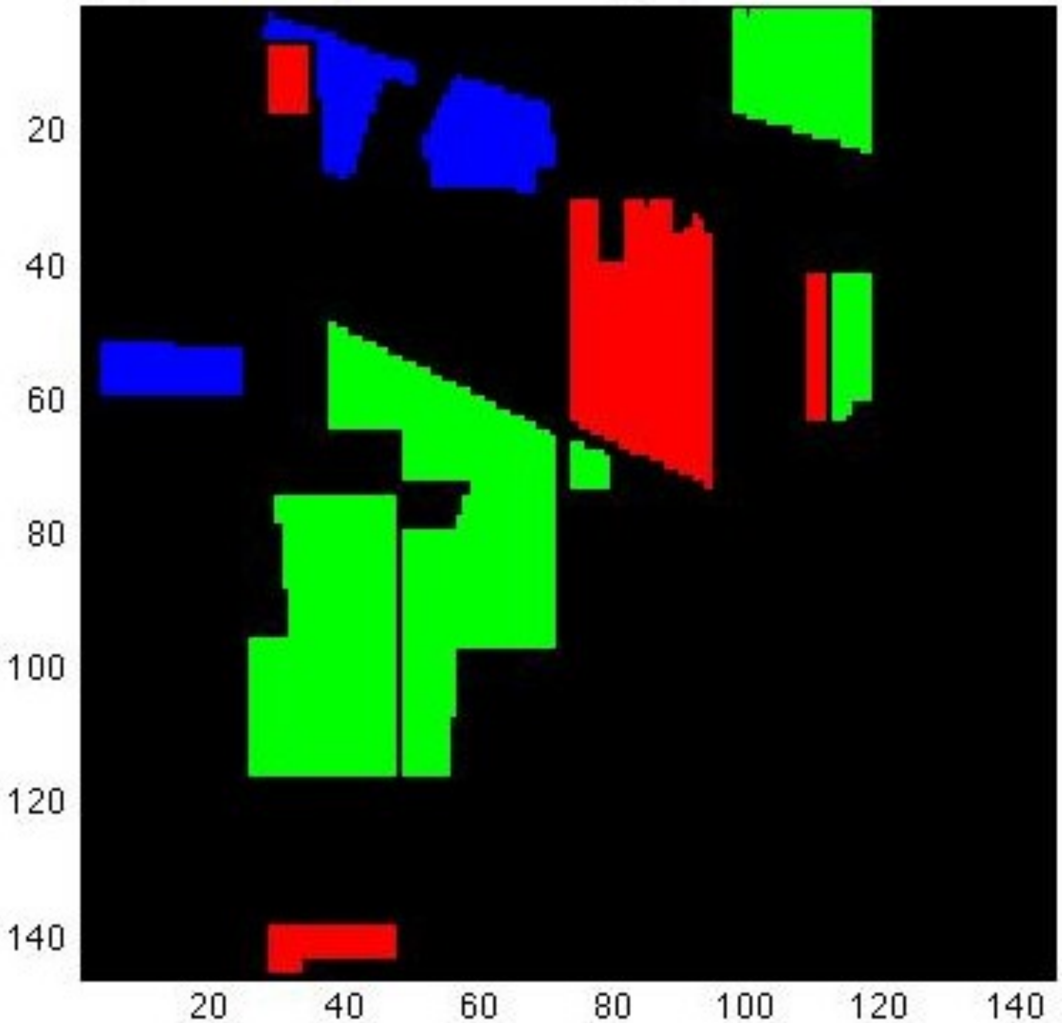
Indian Pines Image





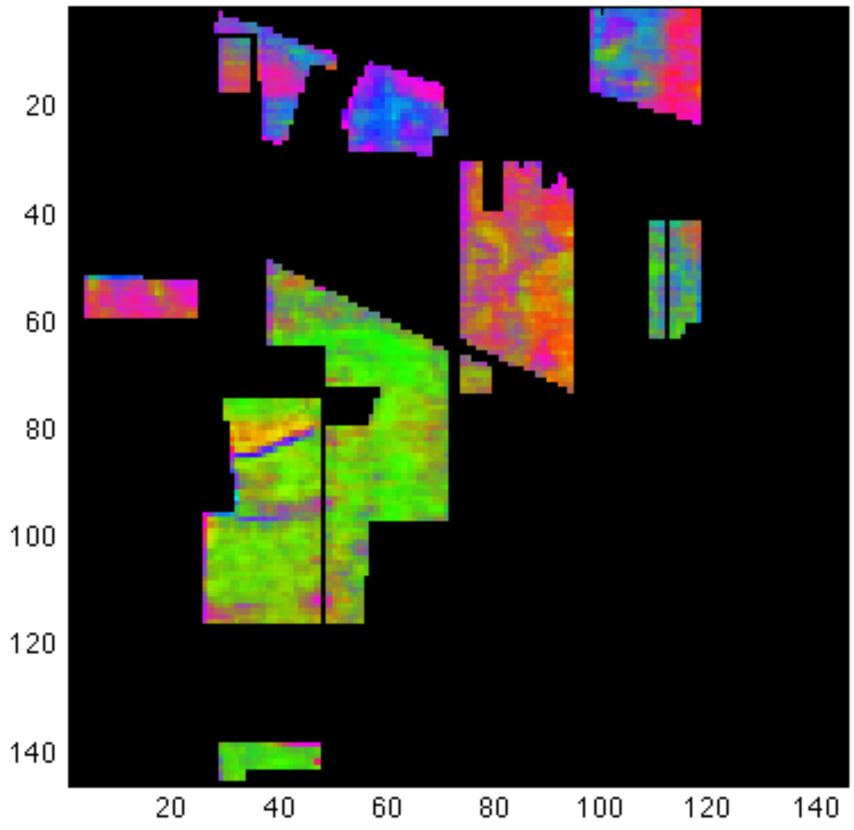
EIGENVECTOR RESEARCH INCORPORATED

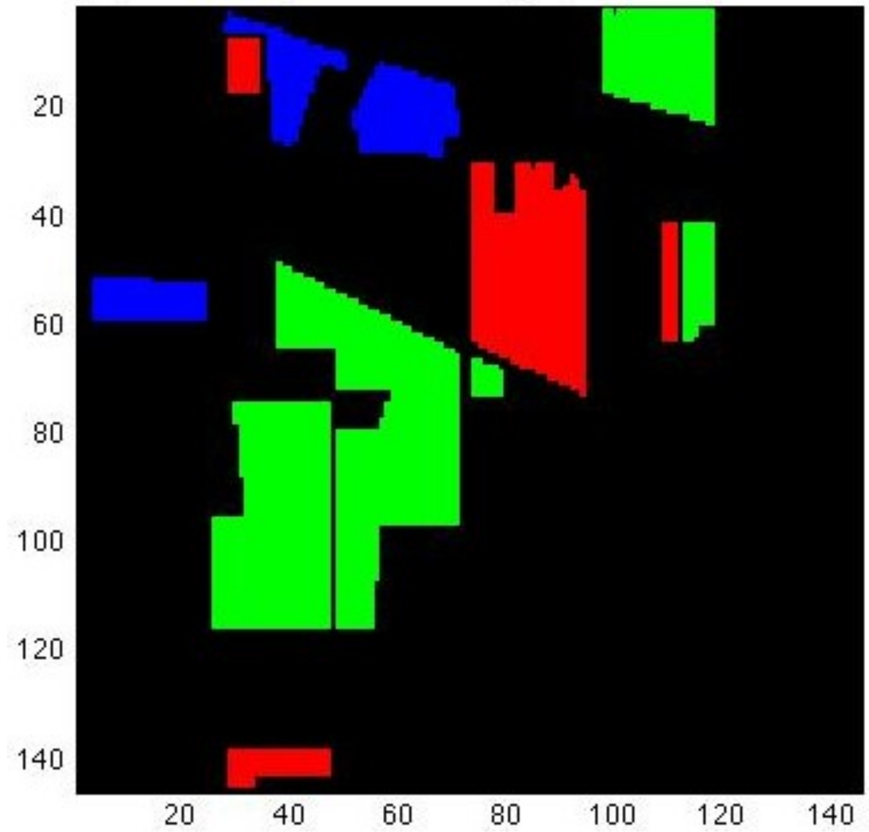
Soybean Fields



Soybeans no till
Soybeans min
Soybeans clean

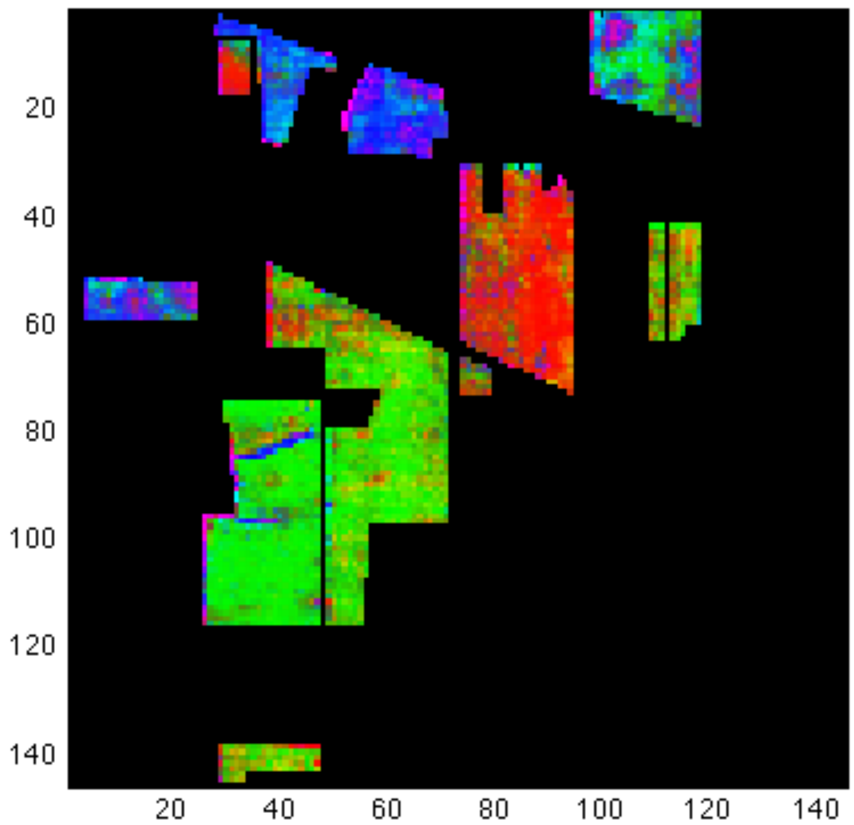
PLS-DA, Mean-Center Only

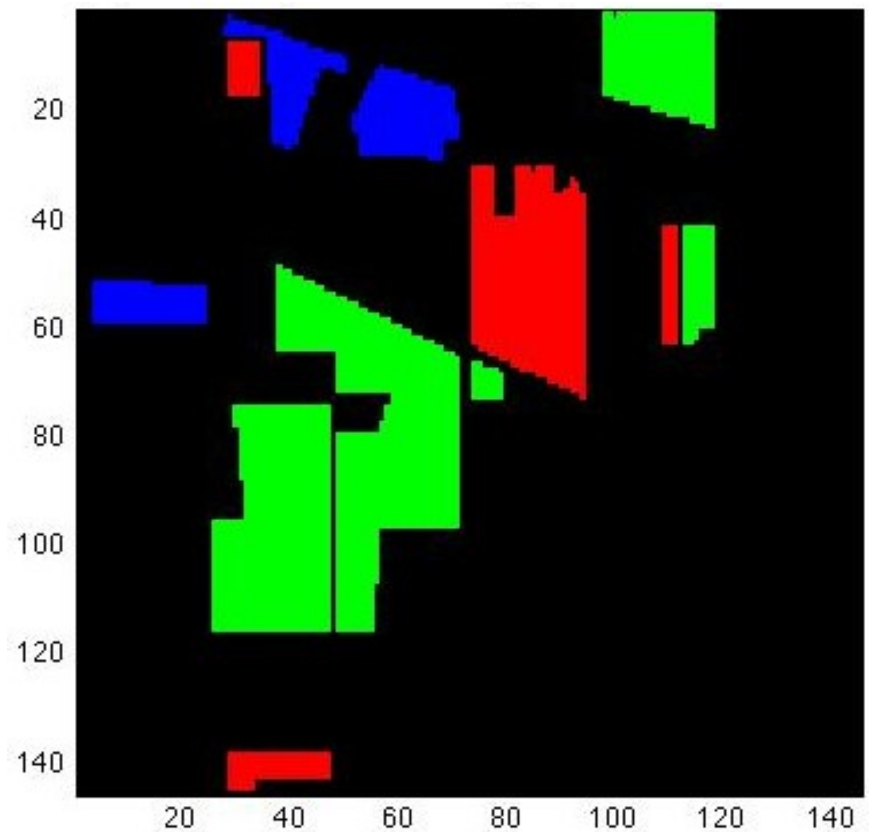




Class Probability Image

PLS-DA, EPO 1-PC



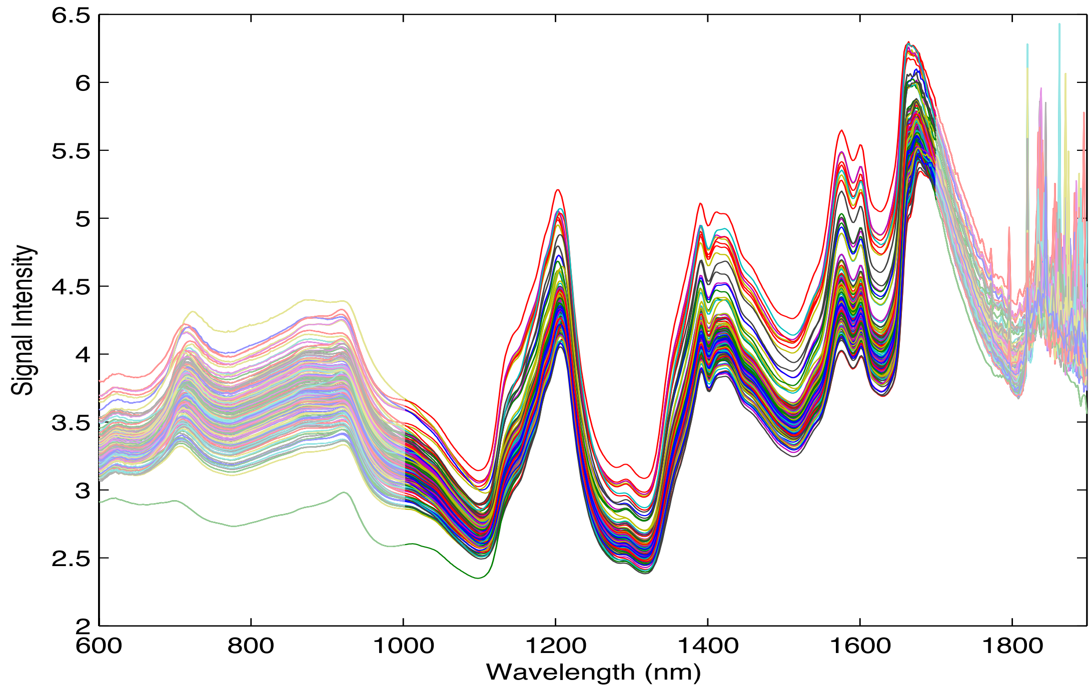


Class Probability Image

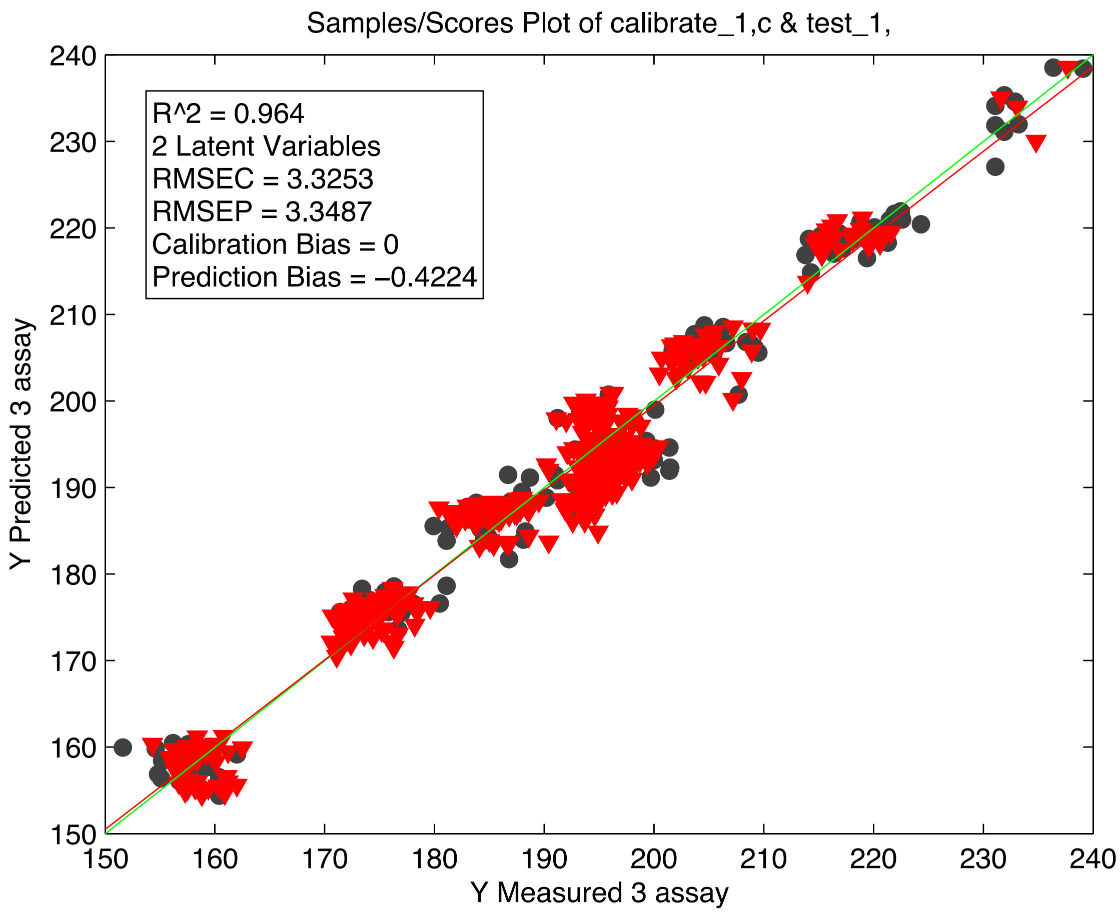
EIGENVECTOR RESEARCH INCORPORATED

Example Calibration Data

- IDRC-2002 Shootout data
- NIR Transflectance of pharmaceutical tablets
- Goal is to predict assay value

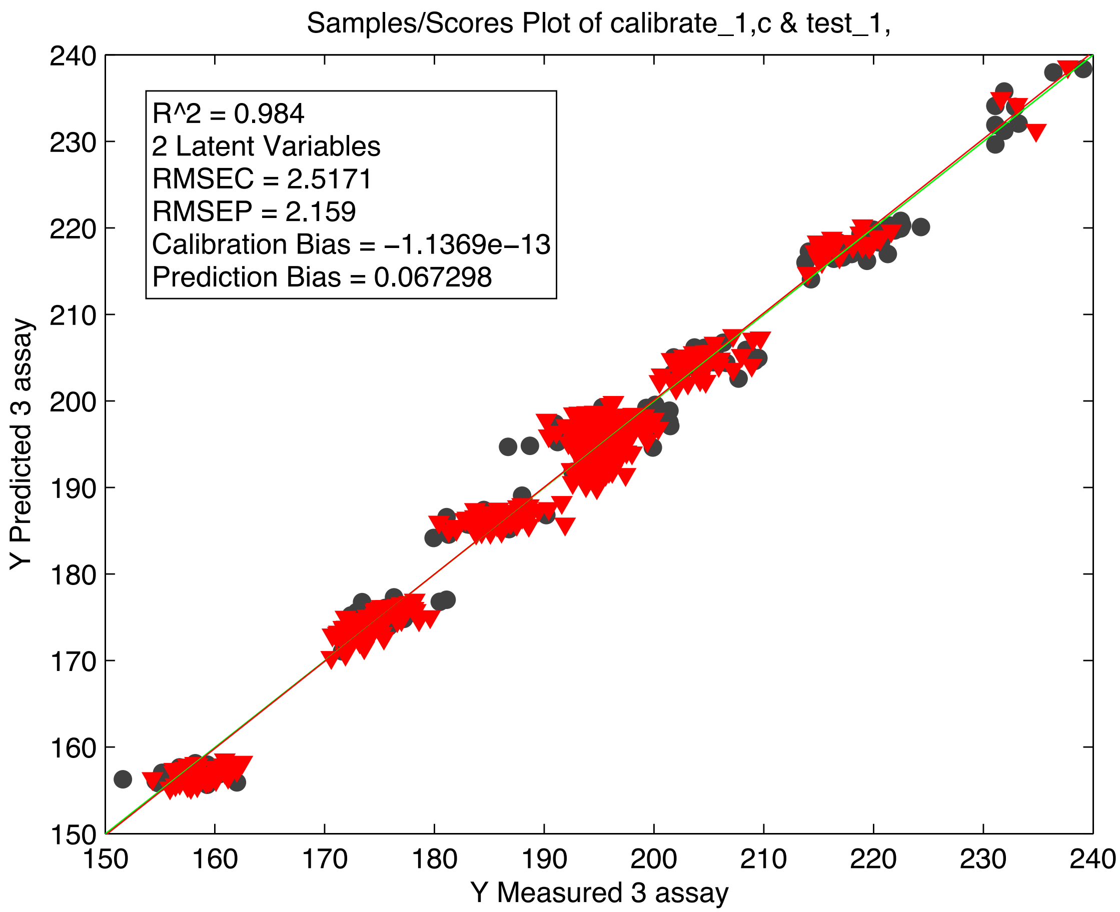


Calibration and Test with MSC & MC

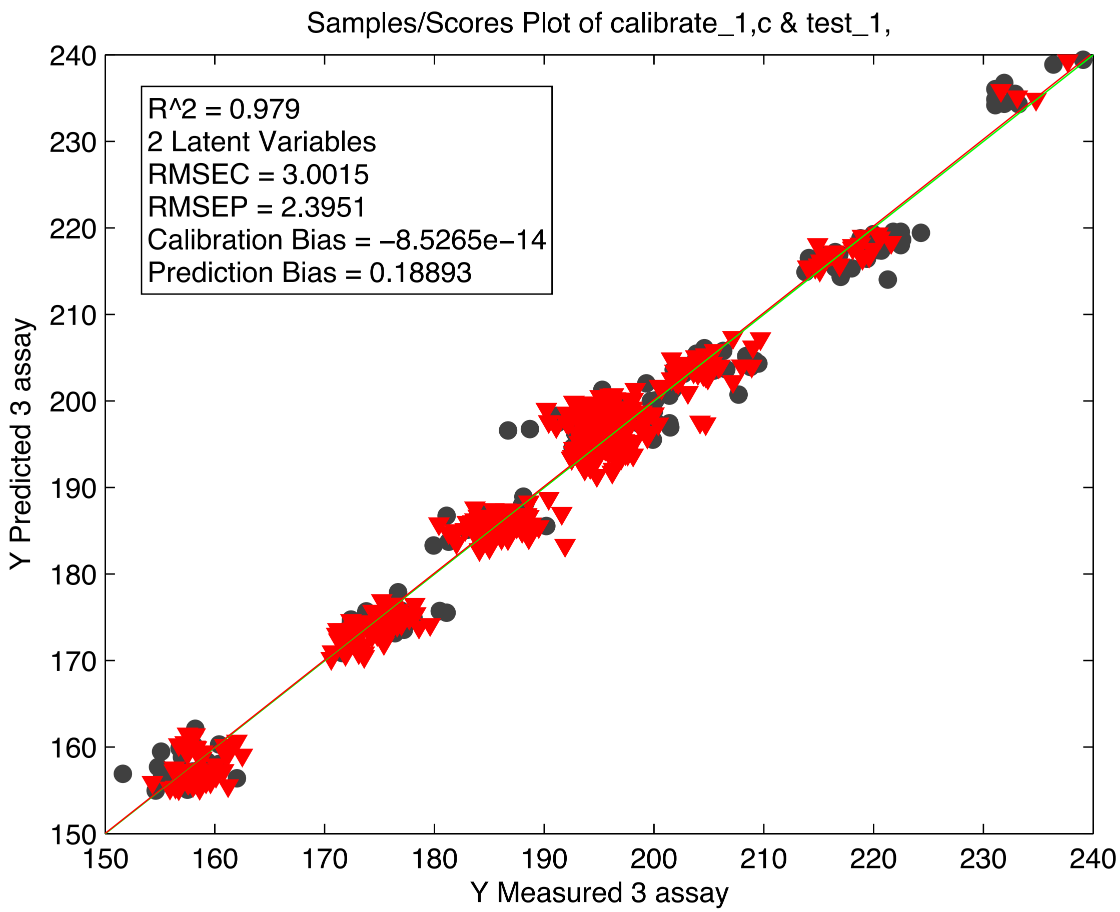


EIGENVECTOR RESEARCH INCORPORATED

With MSC, GLS & MC



With MSC, EPO & MC



EIGENVECTOR RESEARCH INCORPORATED

Orthogonalization Filters

Filter	Soft/ Hard	Adj. Params	Clutter source	Improves Prediction?
OSC	Hard	# LVs	Part of **X** orthogonal to **y**	No, but reduces models complexity
O-PLS	Hard	# LVs	Part of **X**-model space orthogonal to **X'y**	No, but sometimes improves interpretation
MOSC	Hard	# PCs	Part of **X** orthogonal to **y**	Maybe
CPSA	Hard	# PCs	A priori, includes pathlength adj.	Yes
EPO	Hard	# PCs	Classes, y-gradient or a priori	Yes
DOP	Hard	# PCs	Synthetic reference samples	Yes
GLS	Soft	Shrinkage α	Classes, y-gradient or a priori	Yes
SBC	Soft	# PCs (20?)	Repeat samples or blanks	Yes
EMM	Hard	None	A priori from known interferents, clutter subspace	Yes, CLS model
ELS	Hard	# PCs	Clutter subspace	Yes
PA-CLS	Hard	None/# PCs	Baseline shapes, residuals	Yes, CLS model
WLS	Soft	Regularization	Noise measurements	Yes

Conclusions

- Main differences between methods are
  - How the clutter is defined
  - Whether the de-weighting is hard or soft
- Filtering methods are more similar than published statements might have you believe
- Methods achieve similar results, model performance generally improved (except O-PLS, OSC)
- Interpretation of filtered results can be challenging – except OPLS (ideally)

EIGENVECTOR RESEARCH INCORPORATED