Orthogonalization Approaches for Data Preprocessing with Pharmaceutical, Petrochemical and Remote Sensing Applications

Barry M. Wise, Jeremy M. Shaver and

Neal B. Gallagher

Eigenvector Research, Inc.

EIGENVECTOR RESEARCH INCORPORATED

Abstract

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.

Some relationships between these methods are noted, along with ties to older work. Examples are given of the use of the methods in calibration and classification problems in pharmaceutical, petrochemical and remote sensing applications.

EIGENVECTOR RESEARCH INCORPORATED

What is an Orthogonalization Filter?

- Removes spectral patterns from data which are "interfering" with signal of interest
- The interfering species are historically called “clutter” (backgrounds, noise, interferents)
- Filters return spectra with features “removed”
- Weighted subtraction of one or more vectors
- "Soft” orthogonalization is deweighting but not outright complete subtraction

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

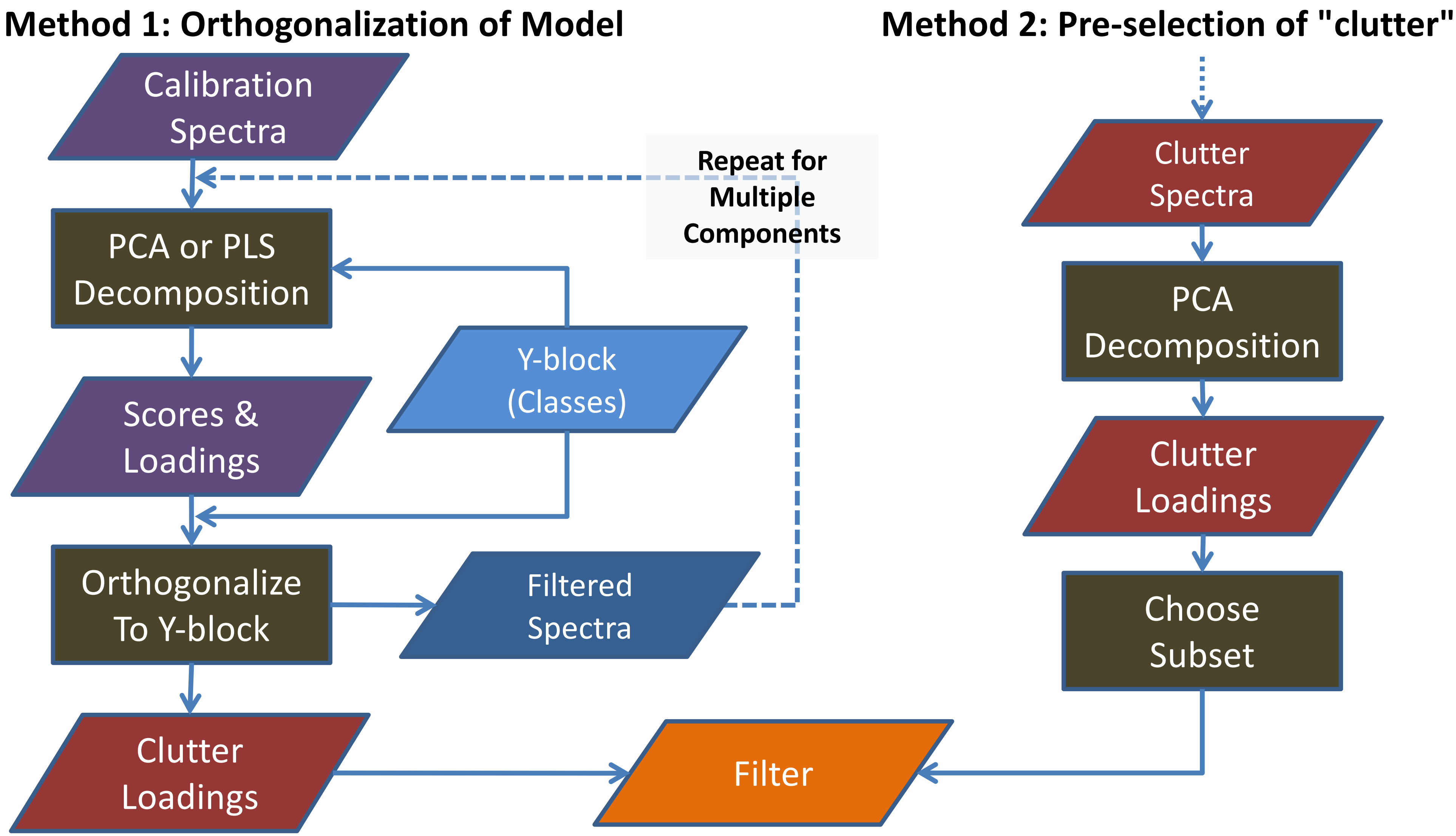
Method 1: Orthogonalization of Model

- 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, Chauchard, Bellon-Maurel 2003)
- GLS – Generalized Least Squares
  (Aitken 1935, Martens et. al. 2003)
- SBC – Science Based Calibration
  (Marbach 2005, patented (?))

Method 2: Pre-selection of "clutter"

EIGENVECTOR RESEARCH INCORPORATED

Two General Approaches



Orthogonal Signal Correction (OSC)

- Introduced by Wold in 1998
  - “OSC paper not the clearest thing” – Johan Trygg, June 9, 2011
- OSC objective function

$$\max_{\mathbf{r}} \left[\mathrm{var}(\mathbf{t}) \mid \mathbf{t} = \mathbf{X}\mathbf{r} \wedge \mathbf{t} \perp \mathbf{y} \right]$$

$$\mathbf{p} = \mathbf{X}^T \mathbf{t} (\mathbf{t}^T \mathbf{t})^{-1}$$

$$\widetilde{\mathbf{X}} = \mathbf{X} - \mathbf{r}^T \mathbf{X}\mathbf{p}$$

EIGENVECTOR RESEARCH INCORPORATED

OSC Issues

- To the extent the objective function is optimized OSC doesn't work
  - Only works if you don't try too hard!
- Many algorithms (at least 5) with various problems
  - Factors not orthogonal to **y**
  - Factors don't capture maximum variance in **X**
  - Filtered **X** not in same subspace as original **X**
- Often implemented prior to cross validation—totally misleading!

O-PLS

- Originally formulated as sequential algorithm (NIPALS based)
- Since shown to be obtainable from post-processing conventional PLS model
- Does not improve prediction
- Claim is that model is more interpretable

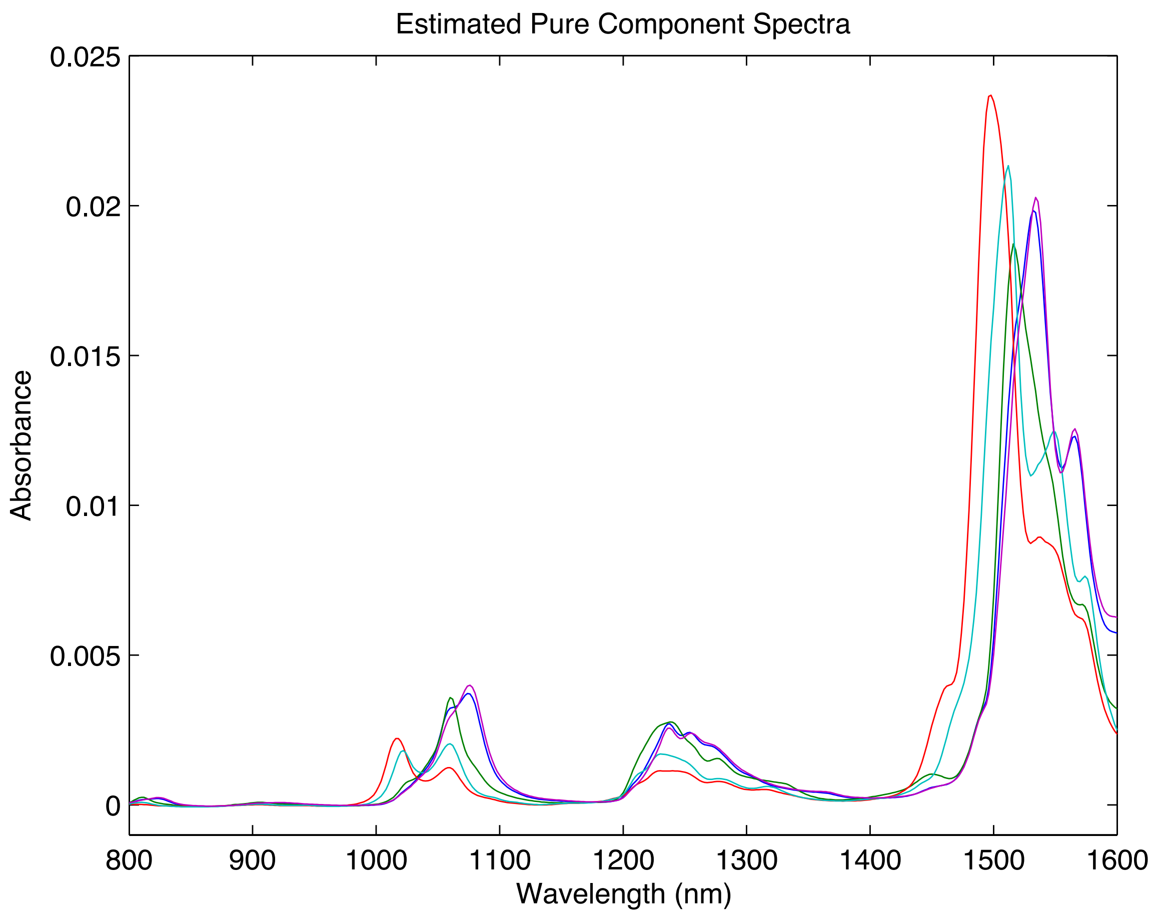
E.K. Kemsley and H.S. Tapp, "OPLS filtered data can be obtained directly from non-orthogonalized PLS1," *J. Chemo*, **23**, 263-264, 2009

R. Ergon, "PLS post-processing by similarity transformation (PLS+ST): a simple alternative to OPLS," *J. Chemo*, **19**, 1-4, 2005

J. Trygg and S. Wold, "Orthogonal Projections to Latent Structures (O-PLS)," *J. Chemo*, **16**, 119-128, 2002.

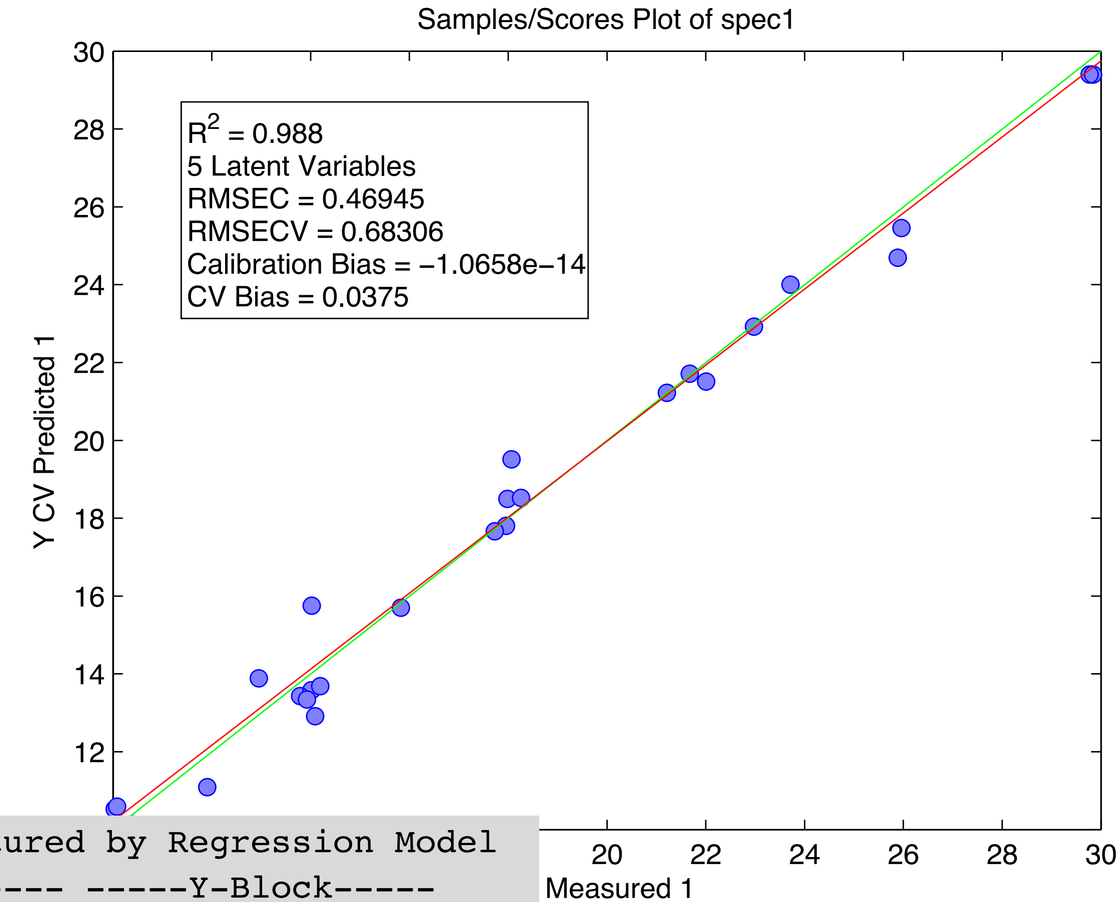
EIGENVECTOR RESEARCH INCORPORATED

NIR of Pseudo-gasoline Samples



EIGENVECTOR RESEARCH INCORPORATED

PLS Model on Component 1

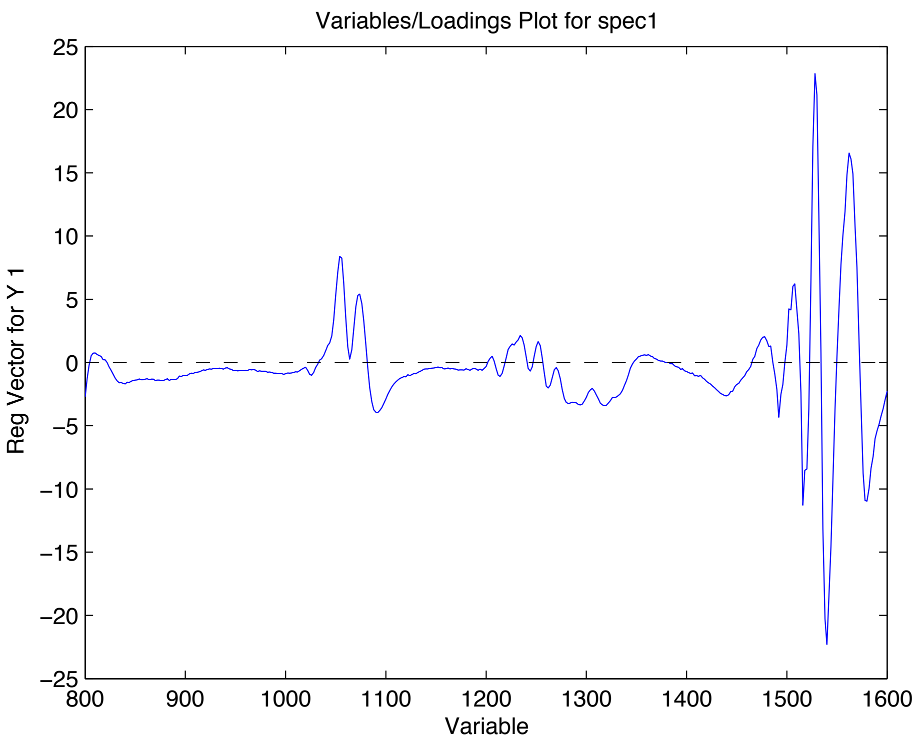


Percent Variance Captured by Regression Model

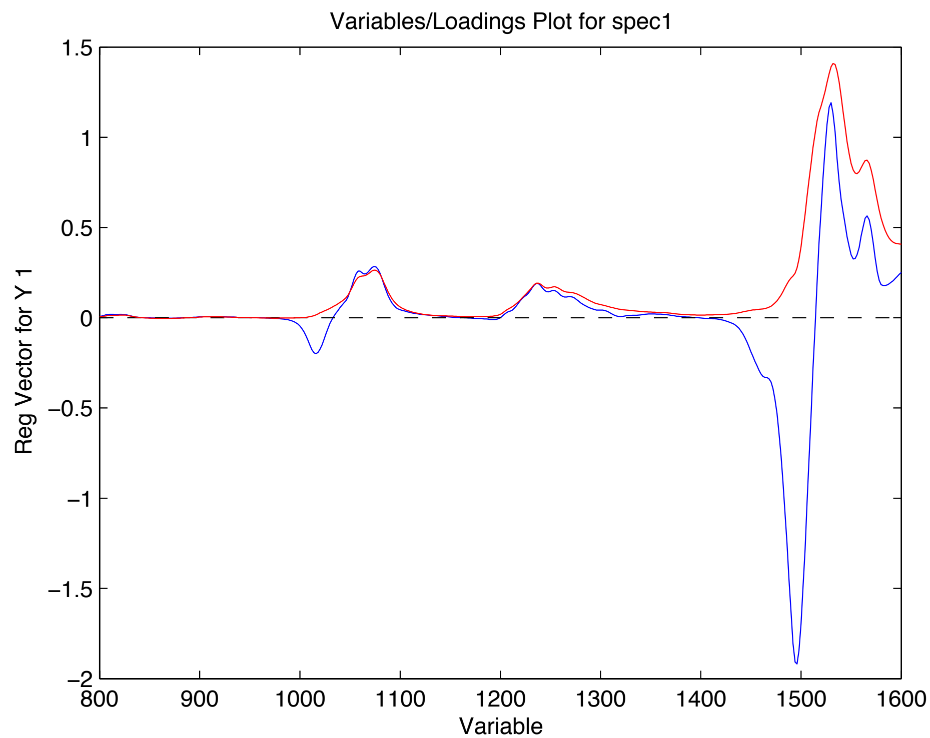
	-----X-Block-----		-----Y-Block-----	
Comp	This	Total	This	Total
1	91.17	91.17	8.36	8.36
2	7.40	98.57	7.19	15.55
3	0.93	99.50	32.81	48.36
4	0.46	99.96	26.18	74.54
5	0.02	99.98	24.90	99.44

EIGENVECTOR RESEARCH INCORPORATED

Regular and O-PLS Filtered Regression Vectors



Regular



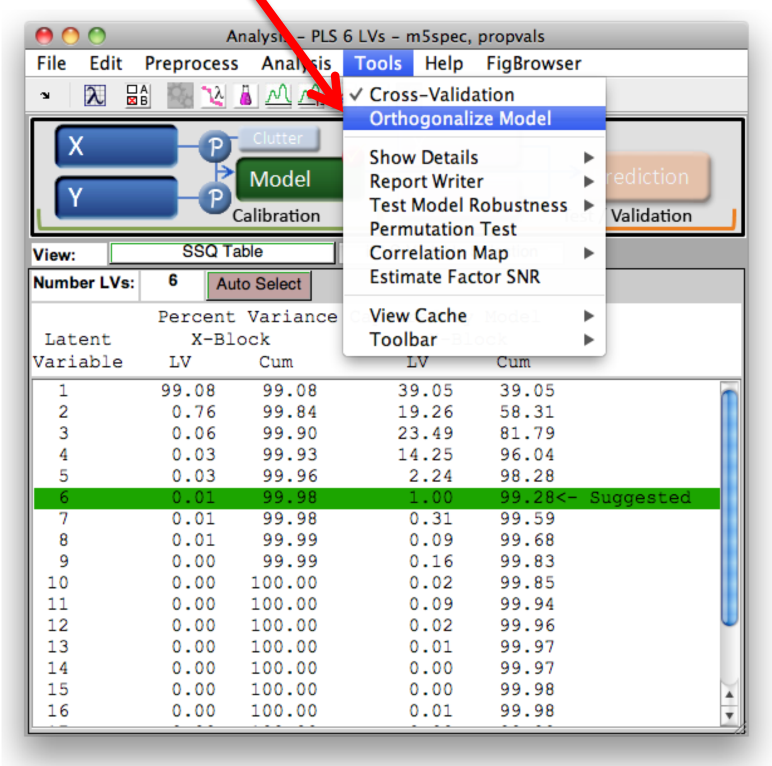
O-PLS Filtered

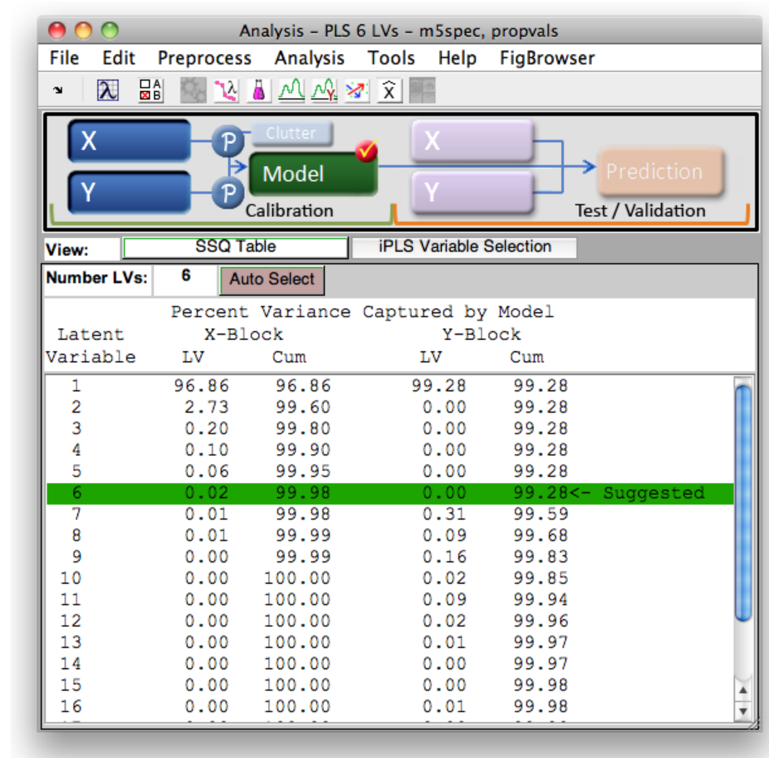
Interpretation

- Better in previous example, but partly because we know what spectra should look like
- What if problem has discrete variables with signal that could be positive, negative or zero?
- Much harder! (see "*On the Interpretability of O-PLS Models*")
- Working on developing better understanding of when it will work and when it won't

EIGENVECTOR RESEARCH INCORPORATED

Orthogonalize Model





EIGENVECTOR
RESEARCH INCORPORATED

Pre-selection Methods…

- CPSA - Constrained Principal Spectral Analysis
  (J. Brown 1990, patented)
- EPO – External Parameter Orthogonalization
  (Roger, Chauchard, Bellon-Maurel 2003)

• Identical
• Choose # of PCs

- GLS – Generalized Least Squares
  (Aitken 1935)
- SBC – Science Based Calibration
  (Marbach 2005, patented (?))

• Quite similar
• Down-weight by scale of eigenvalues

EIGENVECTOR RESEARCH INCORPORATED

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 GLS Weighting Matrix

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

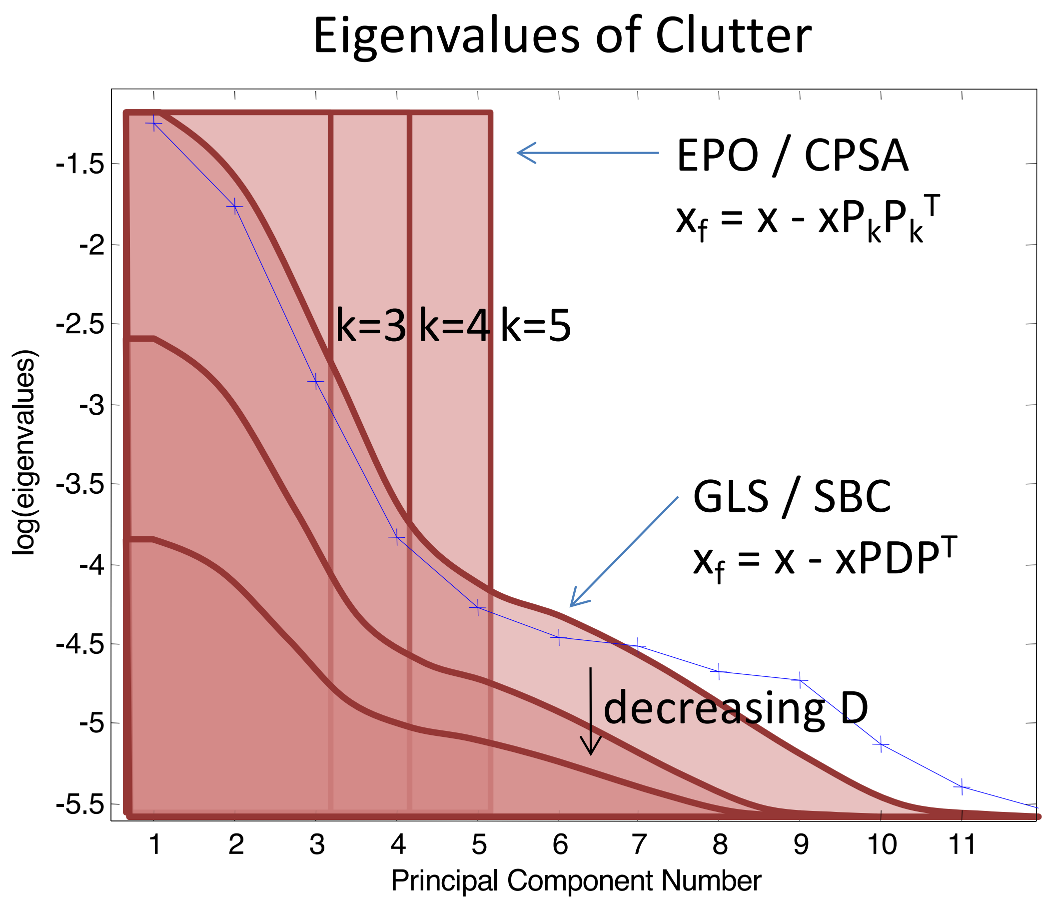
$$\mathbf{G} = \mathbf{V}\mathbf{D}^{-1}\mathbf{V}^{\mathrm{T}}$$

with

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

Large g ➔ 1, dimension unaffected
Small g ➔ 0, dimension eliminated

Choosing Components



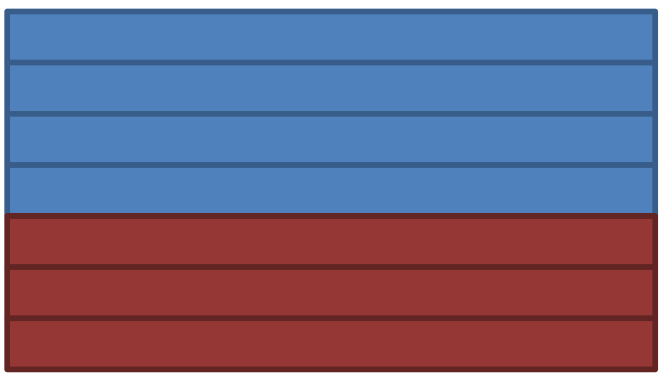
One adjustable parameter in each method

EIGENVECTOR RESEARCH INCORPORATED

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

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

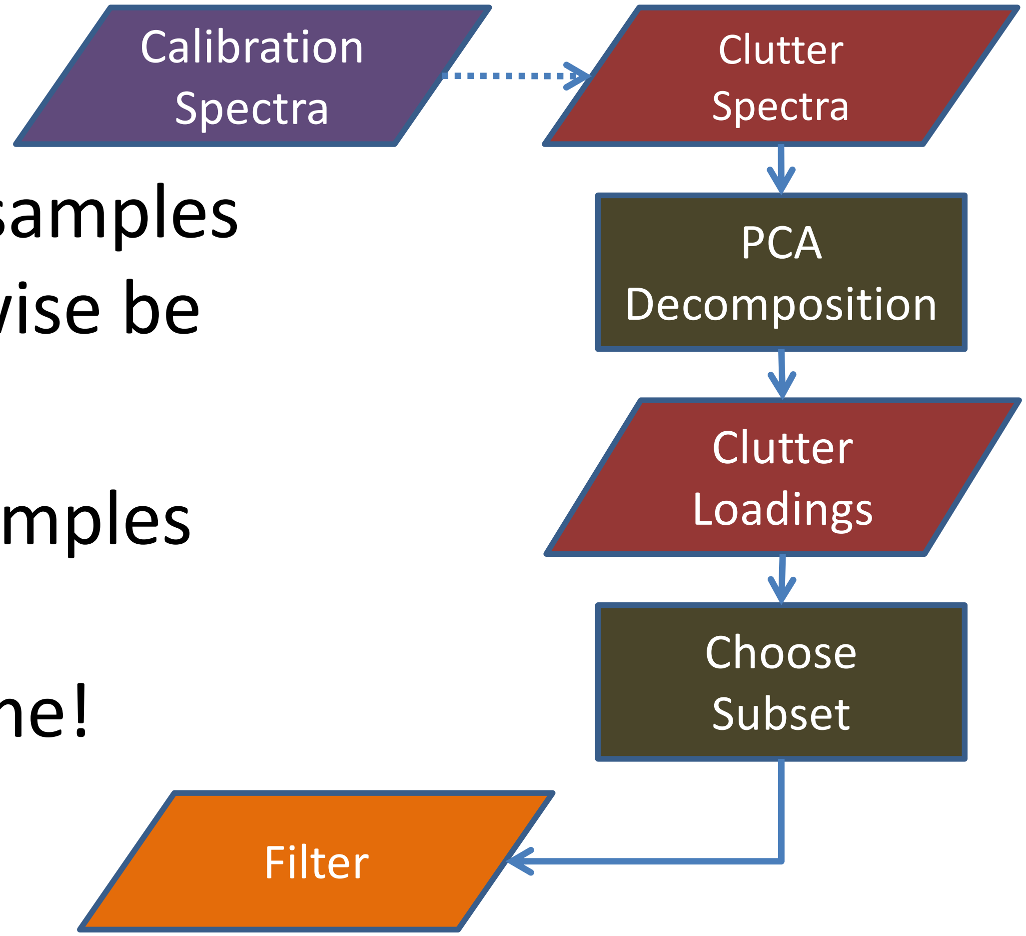
Pseudo-inverse is an orthogonalization!

Equivalent to full-rank EPO / CPSA model

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!



EIGENVECTOR RESEARCH INCORPORATED

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
  - differences between samples where analyte of interest is (nearly) the same, *e.g.* y-gradient
  - repeat measurement of blanks
- Make it up! *e.g.* polynomial baseline shapes

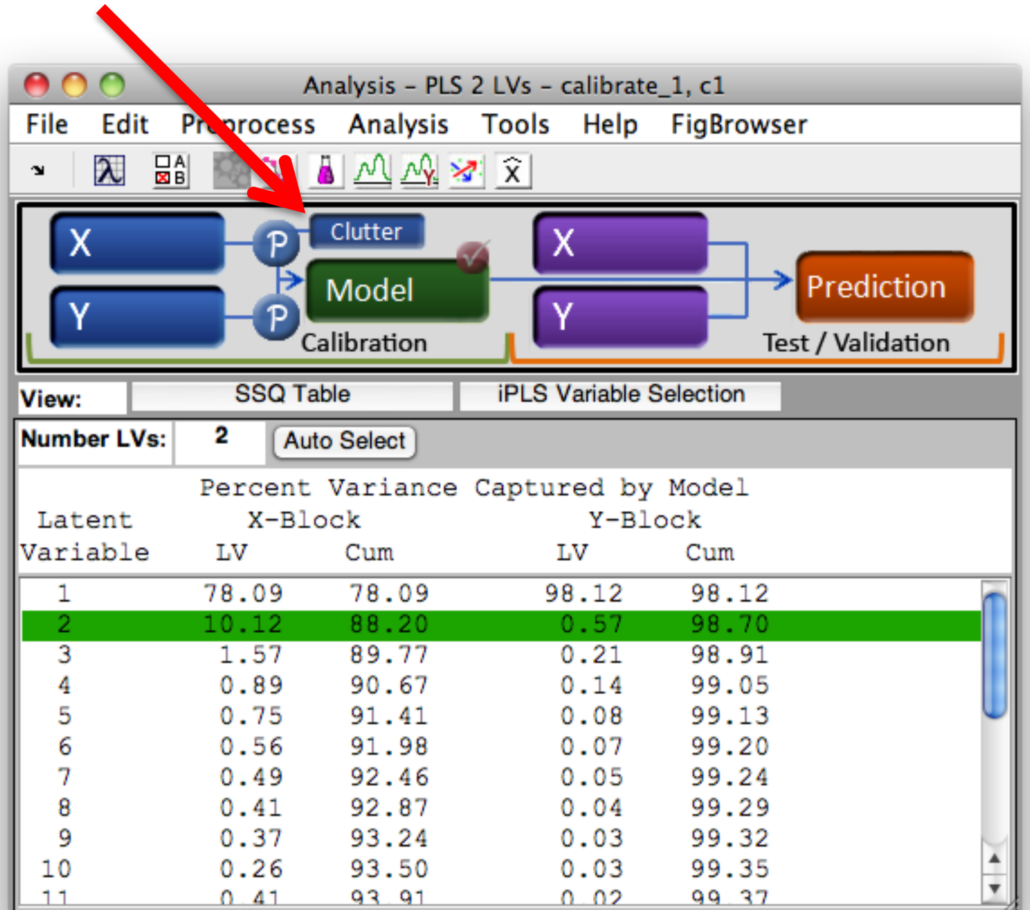
Y-gradient Method

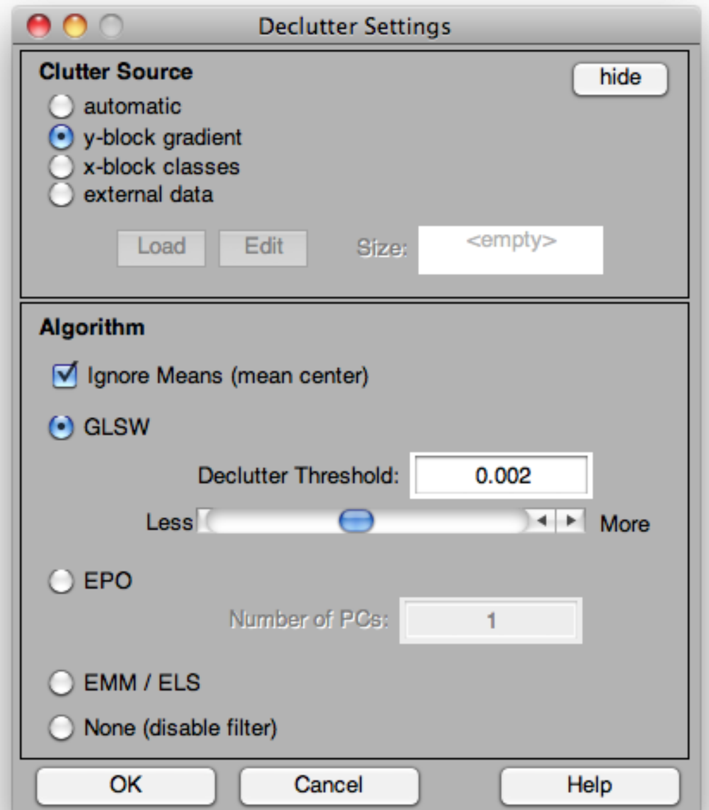
- Sort samples by **y** (reference) values
- Take differences between adjacent samples
- Weight **X**-differences by inverse of difference in **y** values
- Deweight by covariance of differences (GLS) or orthogonalize against some number of PCs (EPO, ELS, EMM, PA-CLS)

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 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

We think it is useful to use Clutter!

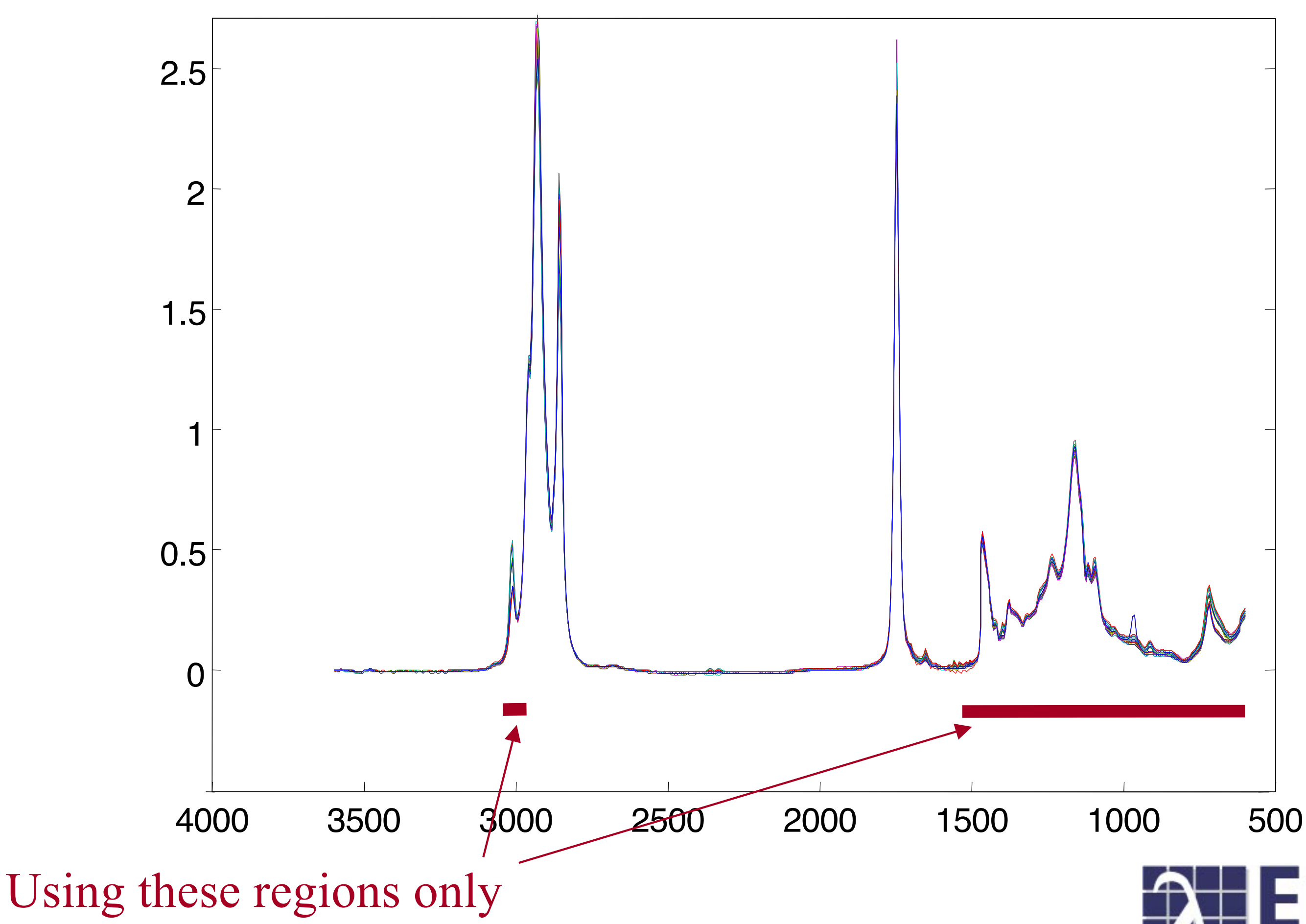




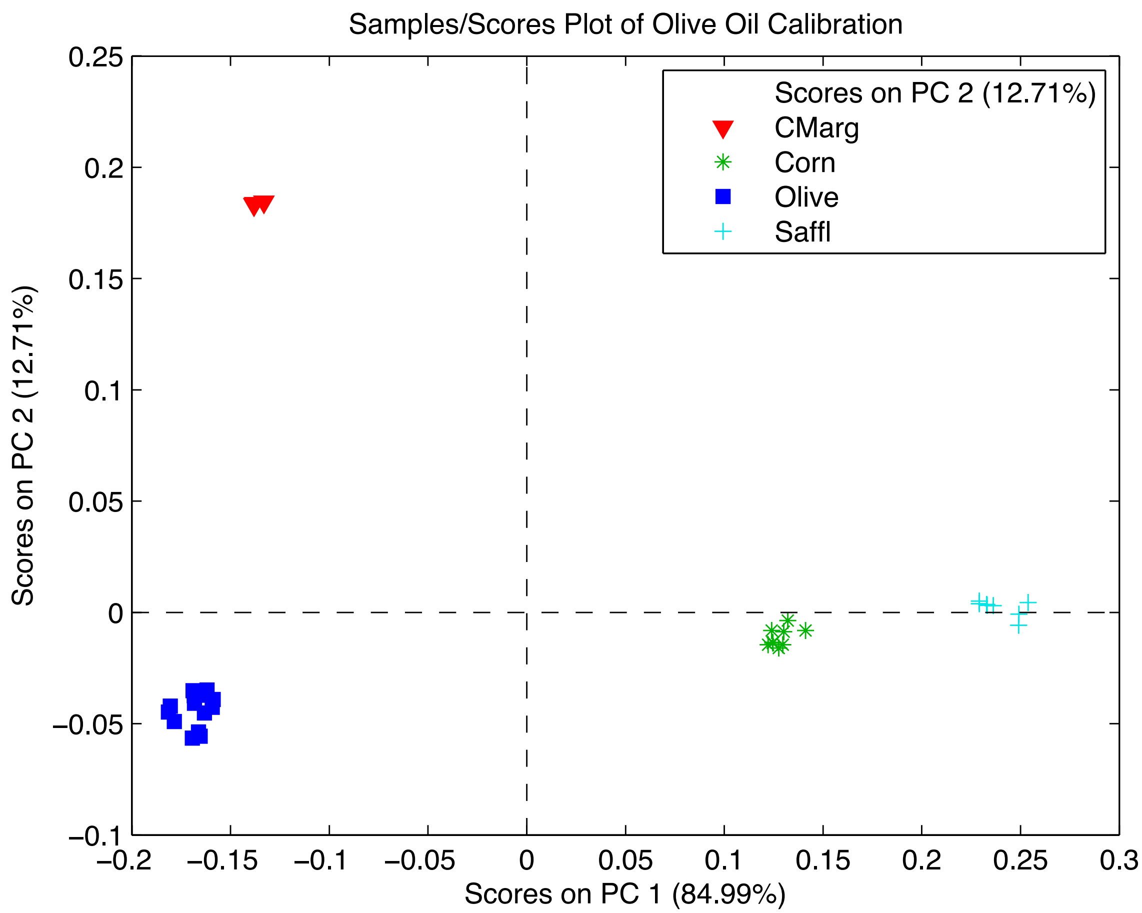
EIGENVECTOR
RESEARCH INCORPORATED

Example Classification Data

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

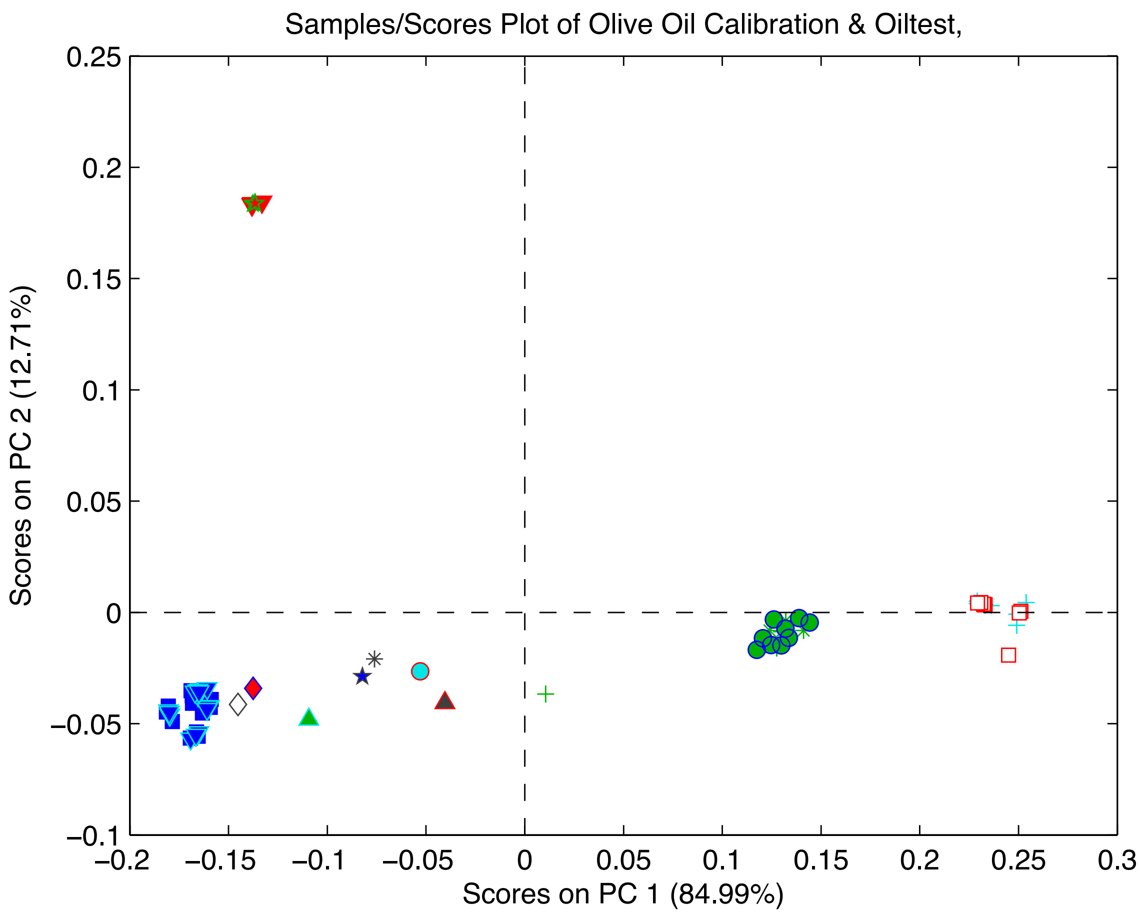


Calibration with MSC

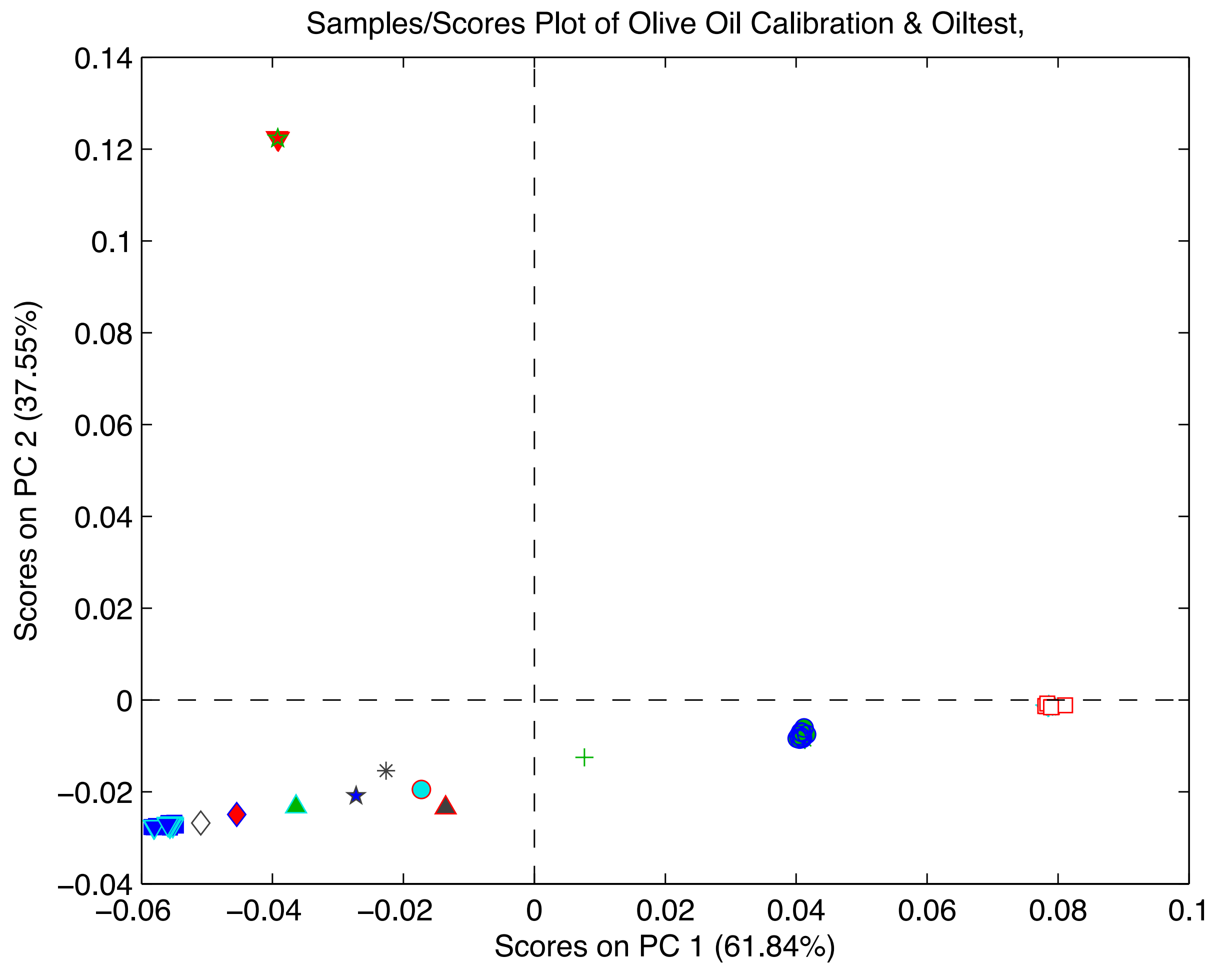


EIGENVECTOR RESEARCH INCORPORATED

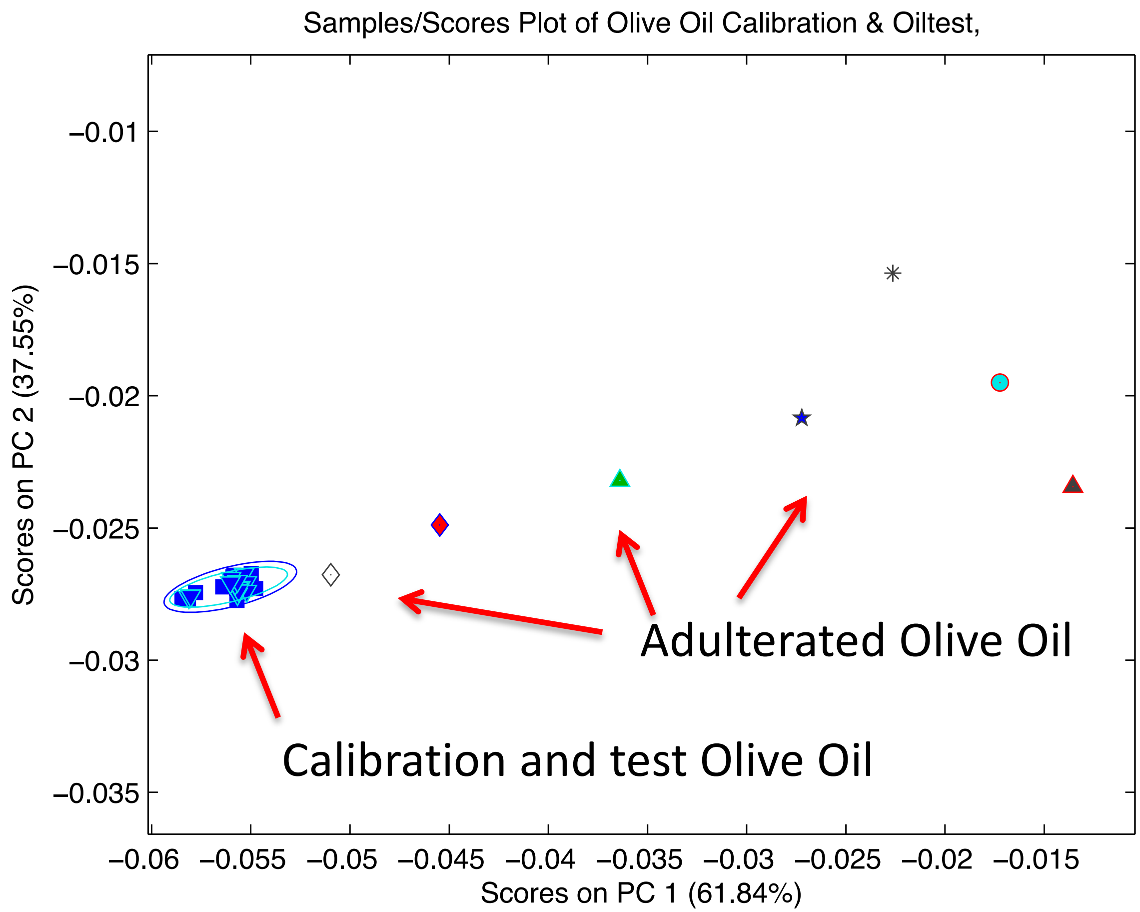
Cal and Test with MSC



With MSC and GLS

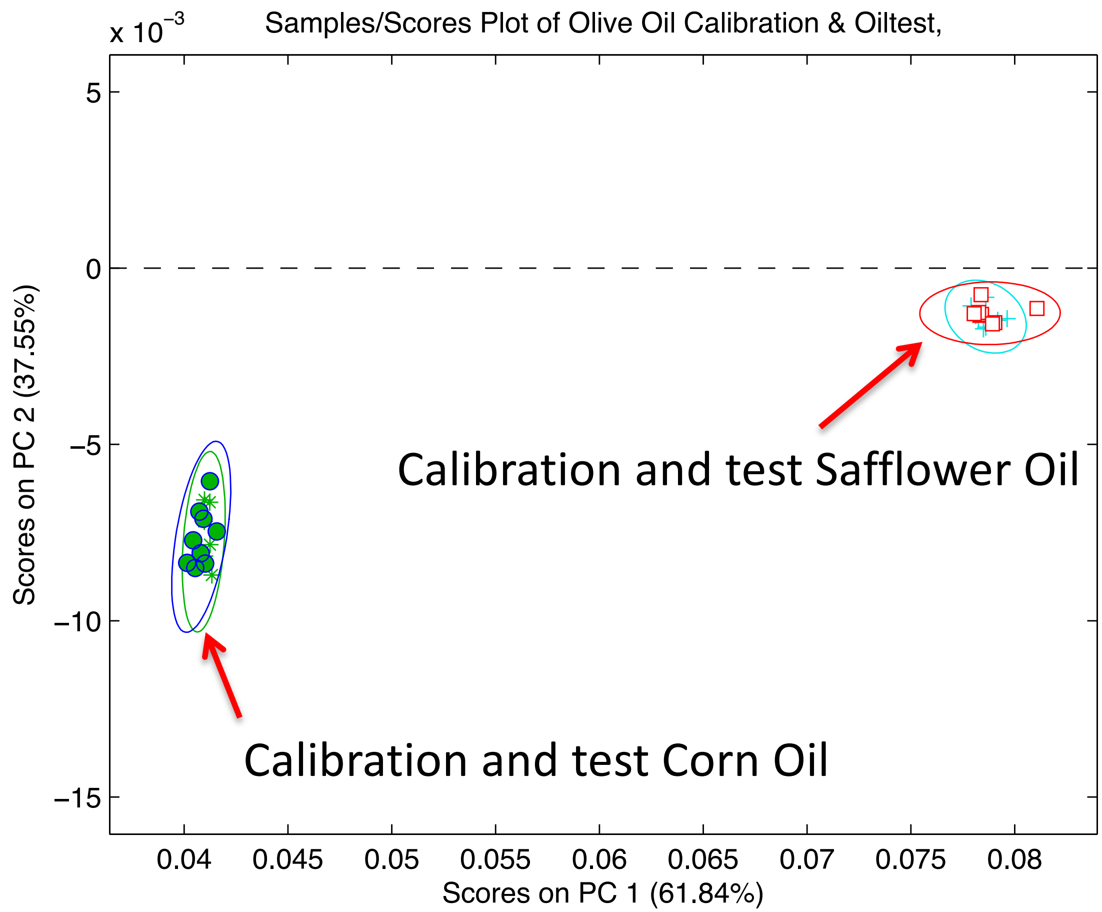


Zoom on Olive Oil



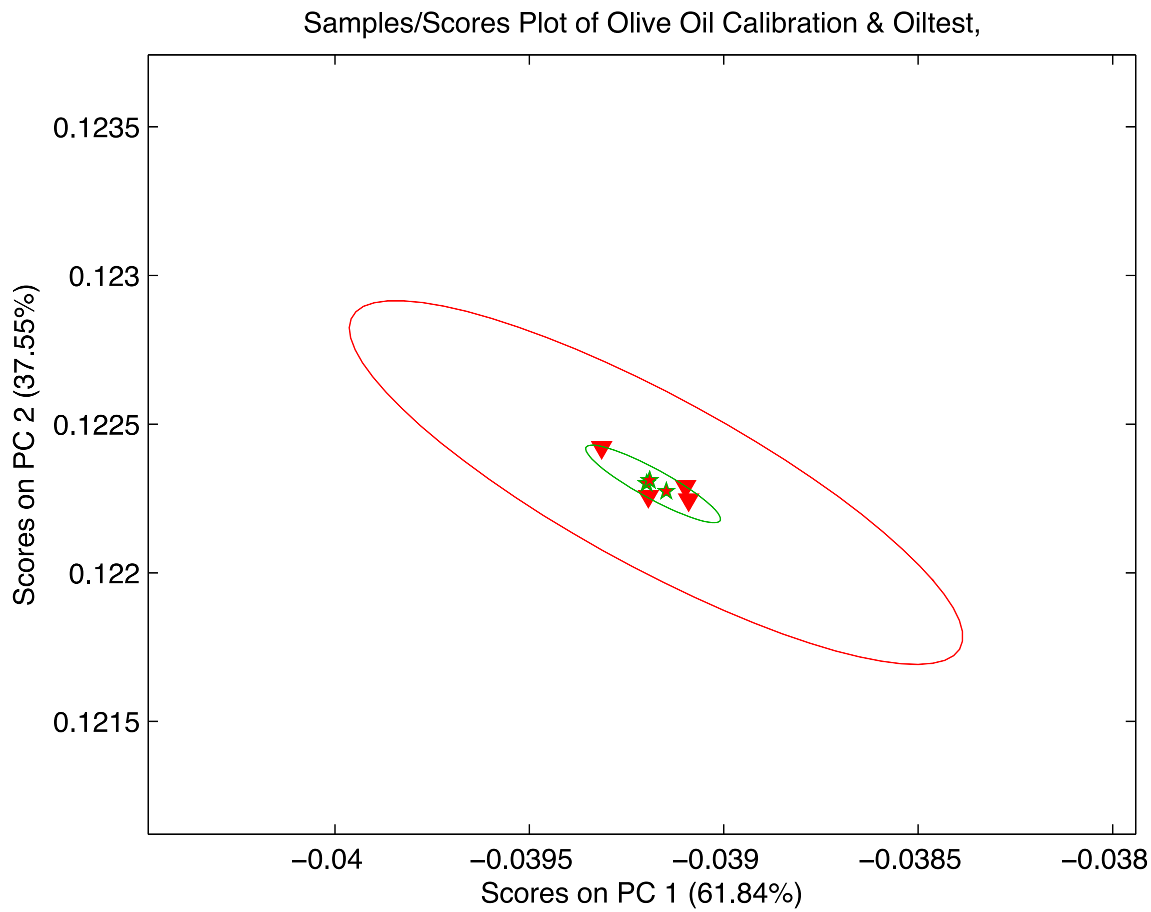
EIGENVECTOR RESEARCH INCORPORATED

Zoom on Corn and Safflower Oil



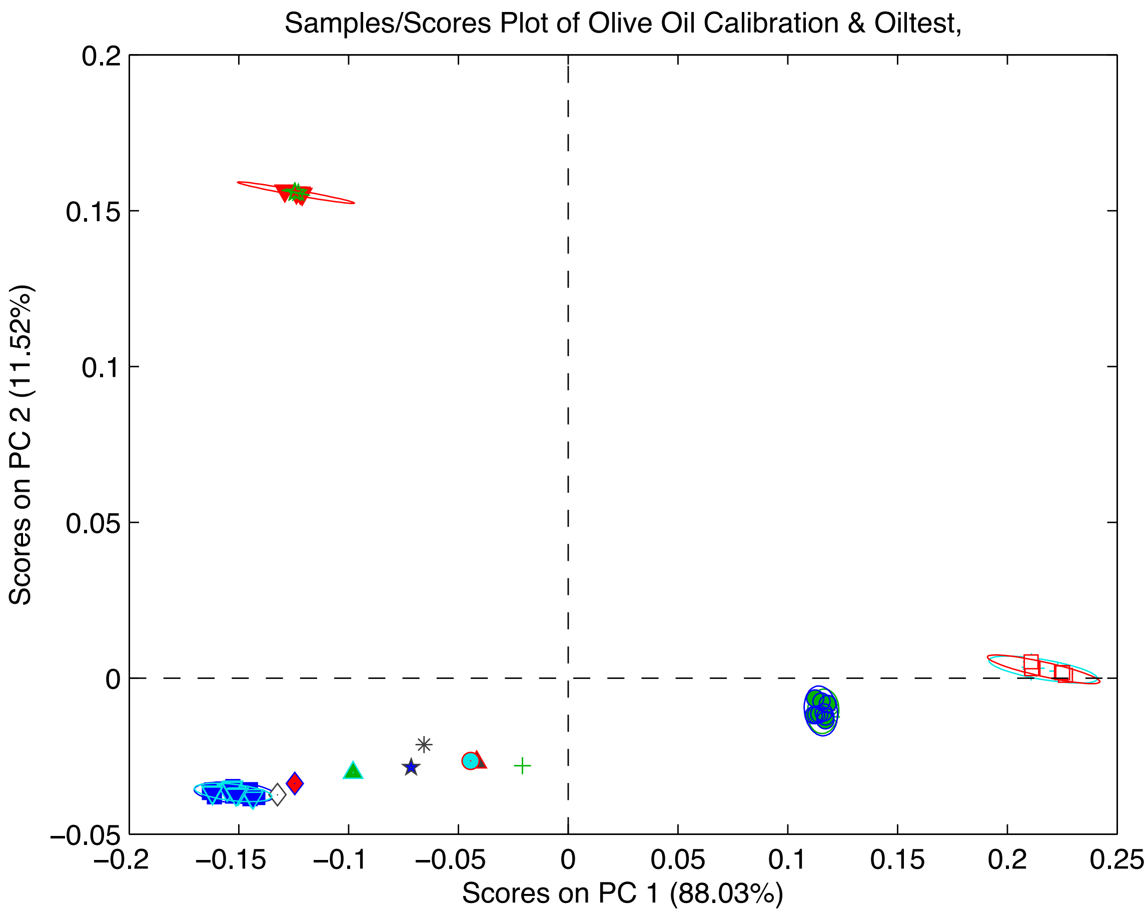
EIGENVECTOR RESEARCH INCORPORATED

Zoom on Corn Margarine



EIGENVECTOR RESEARCH INCORPORATED

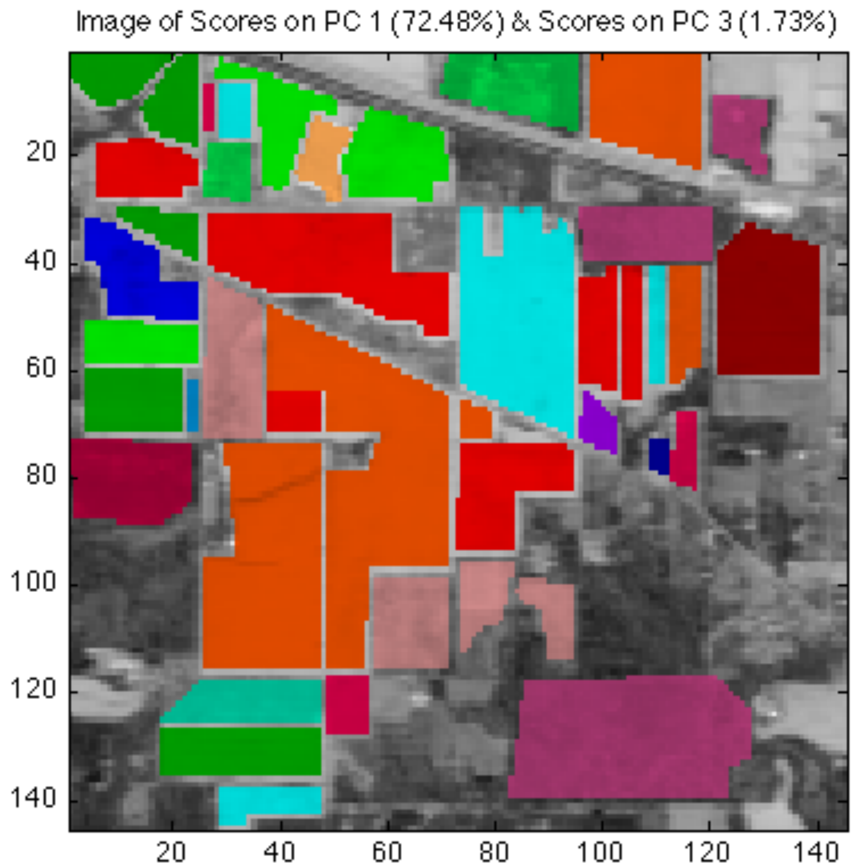
With MSC and EPO

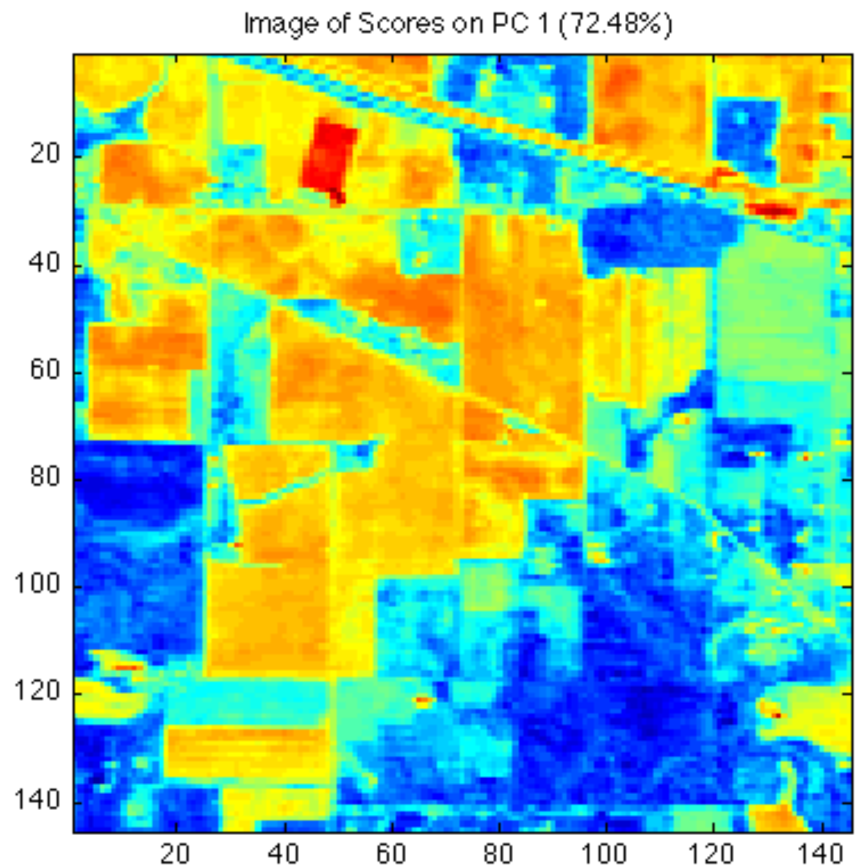


Indian Pines Data

- 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

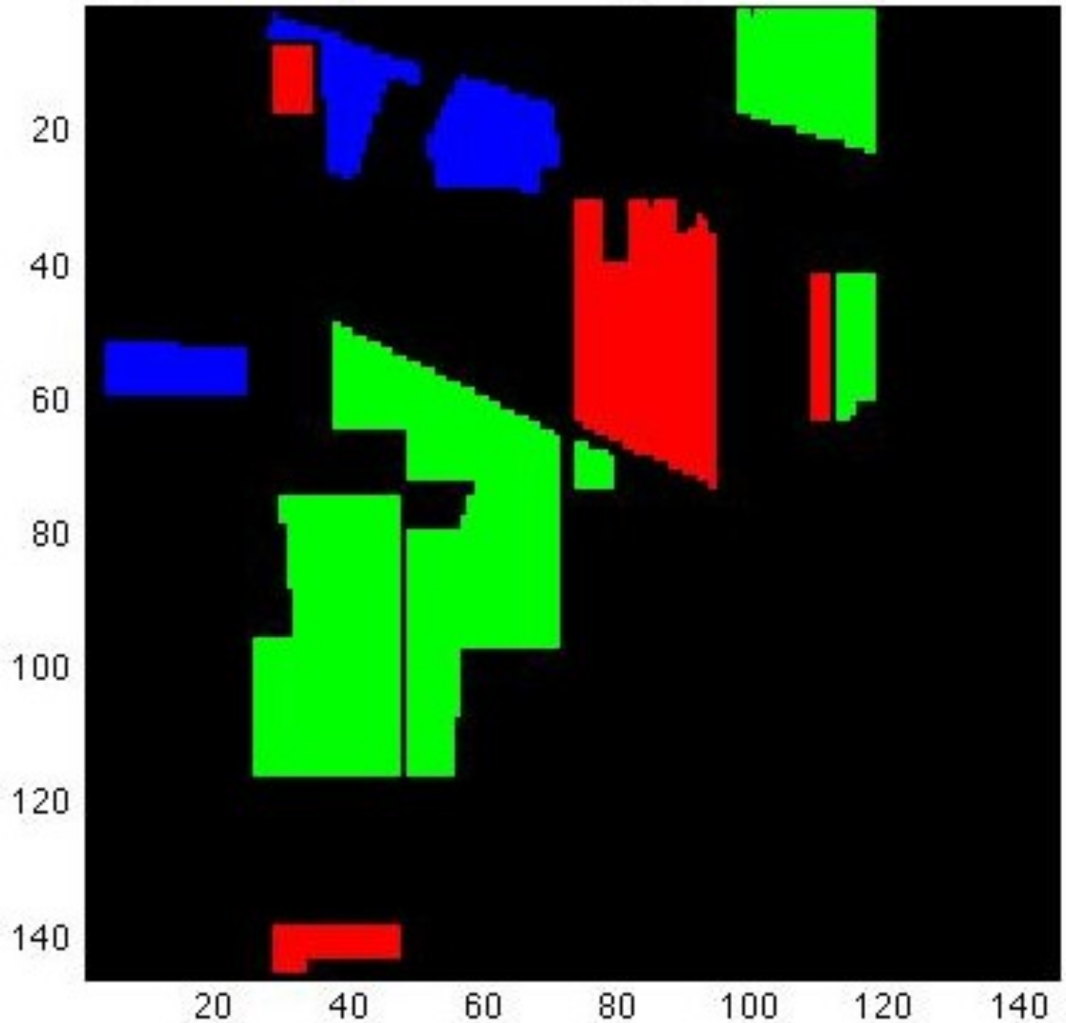
Indian Pines Image





EIGENVECTOR
RESEARCH INCORPORATED

Soybean Fields



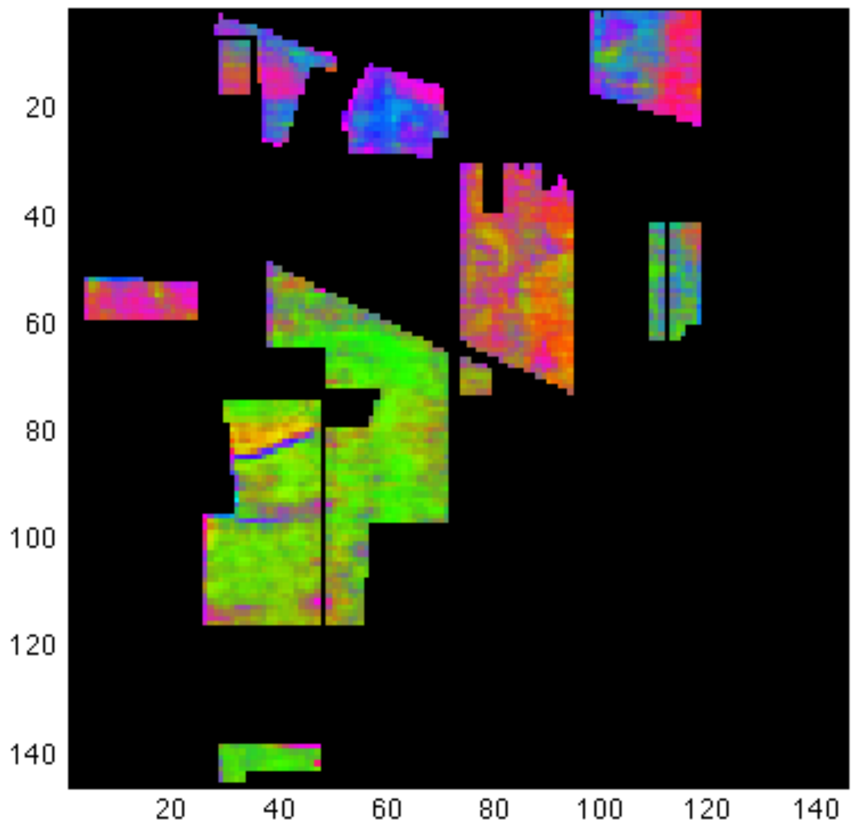
Soybeans no till

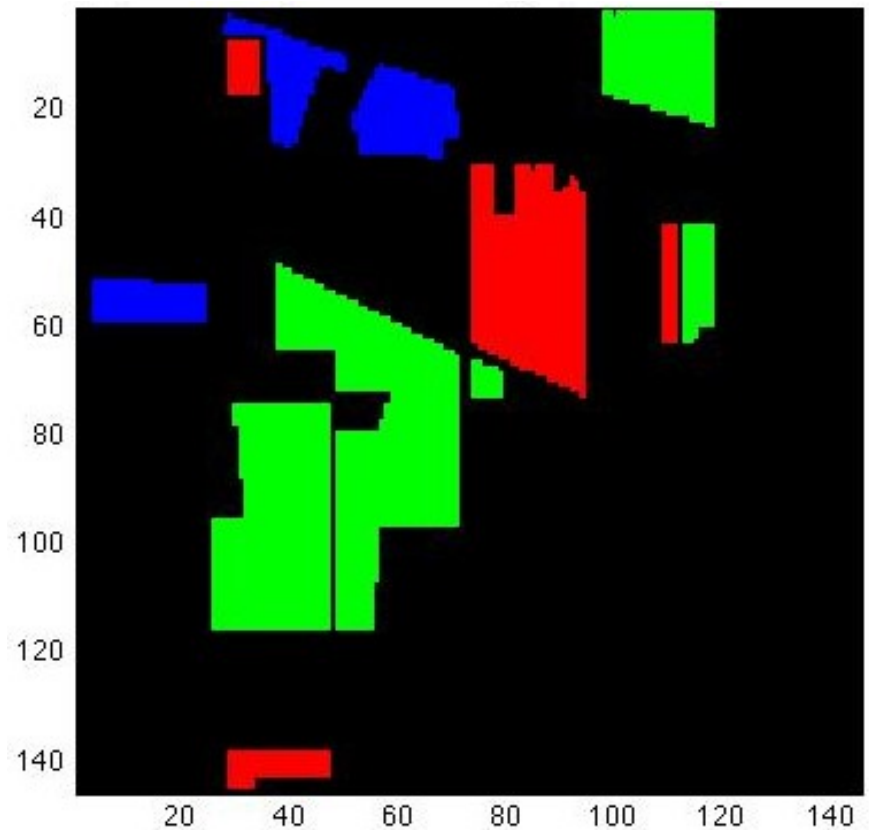
Soybeans min

Soybeans clean

EIGENVECTOR RESEARCH INCORPORATED

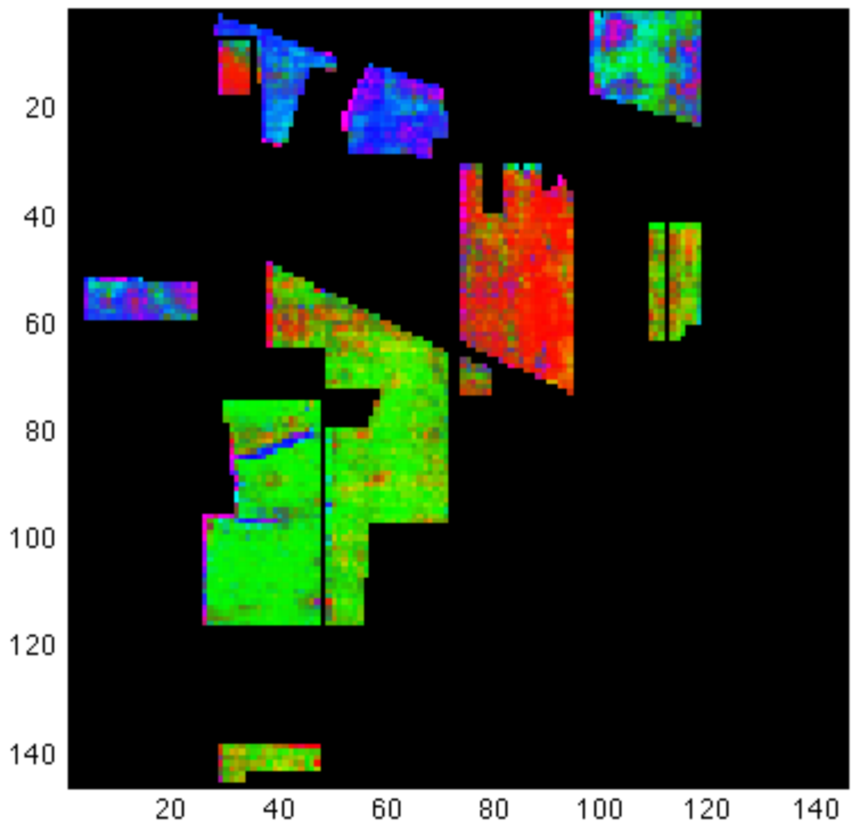
PLS-DA, Mean-Center Only

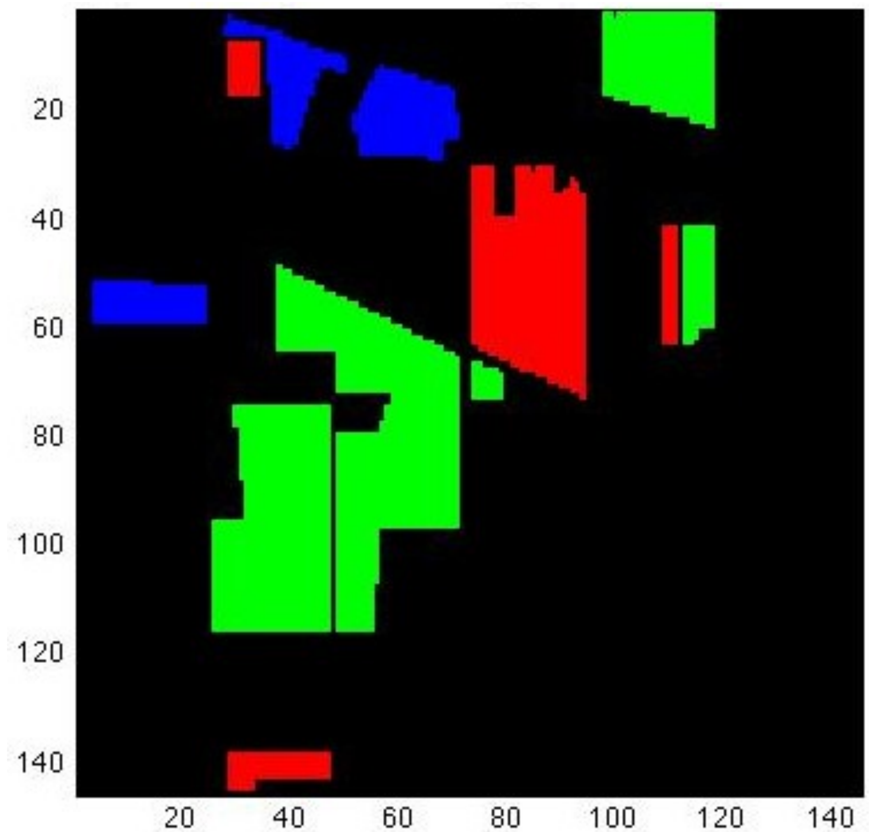




Class Probability Image

PLS-DA, EPO 1-PC



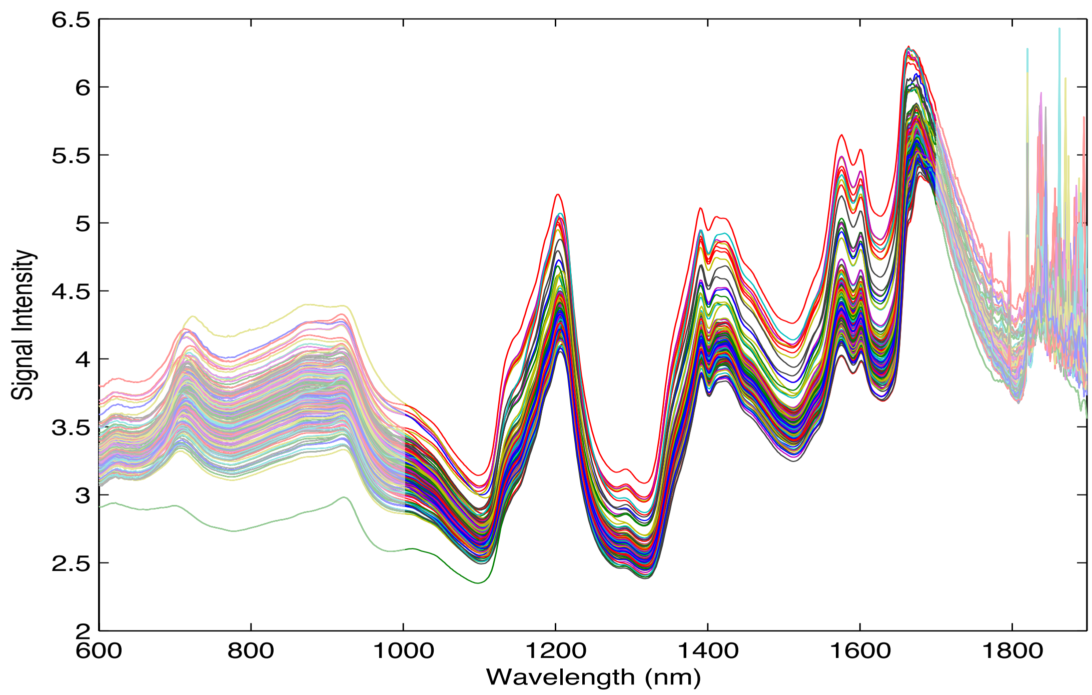


Class Probability Image

EIGENVECTOR RESEARCH INCORPORATED

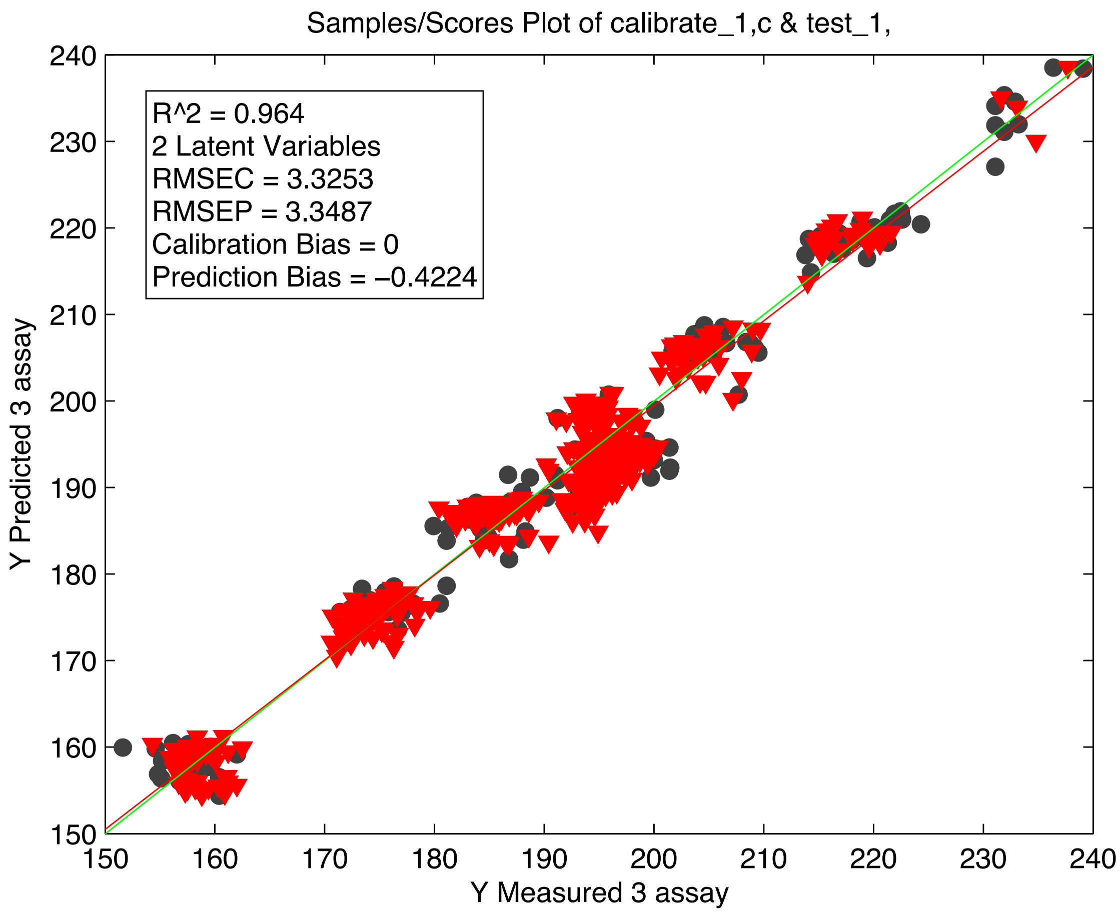
Example Calibration Data

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

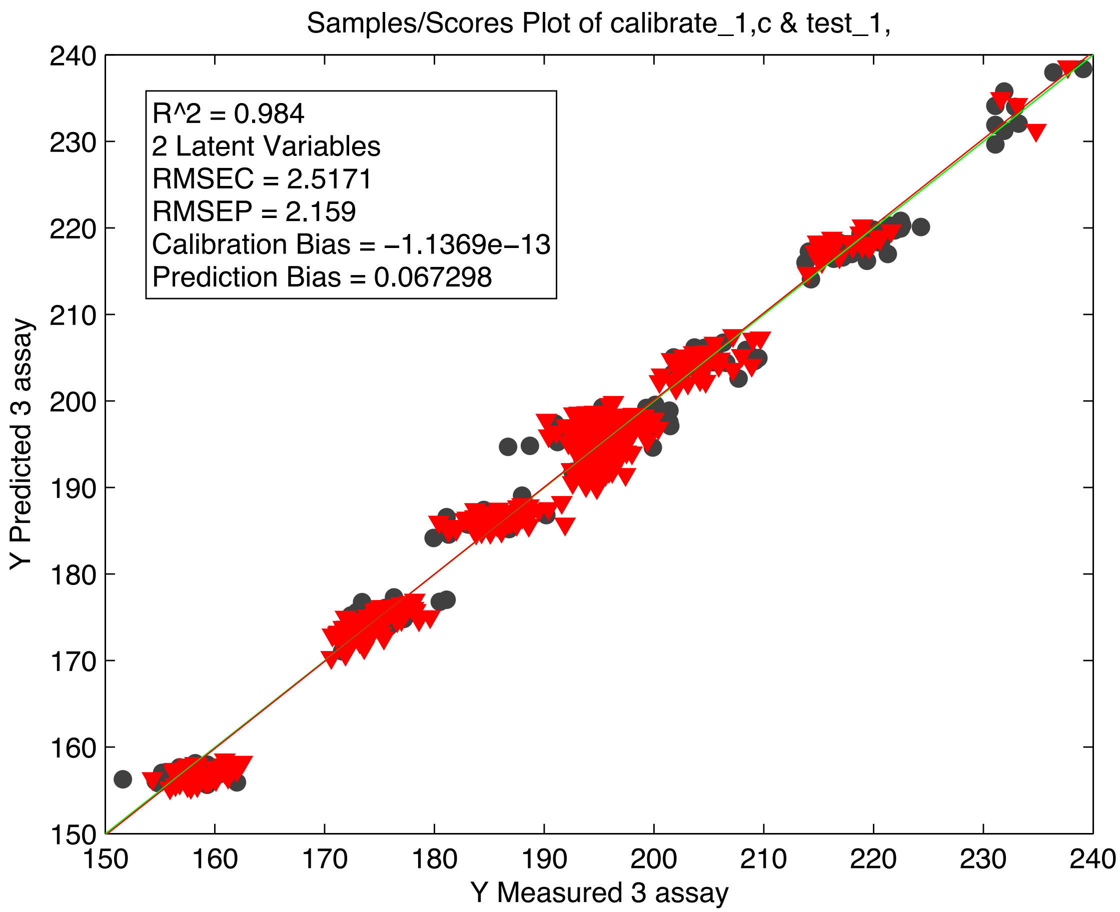


EIGENVECTOR RESEARCH INCORPORATED

Calibration and Test with MSC & MC

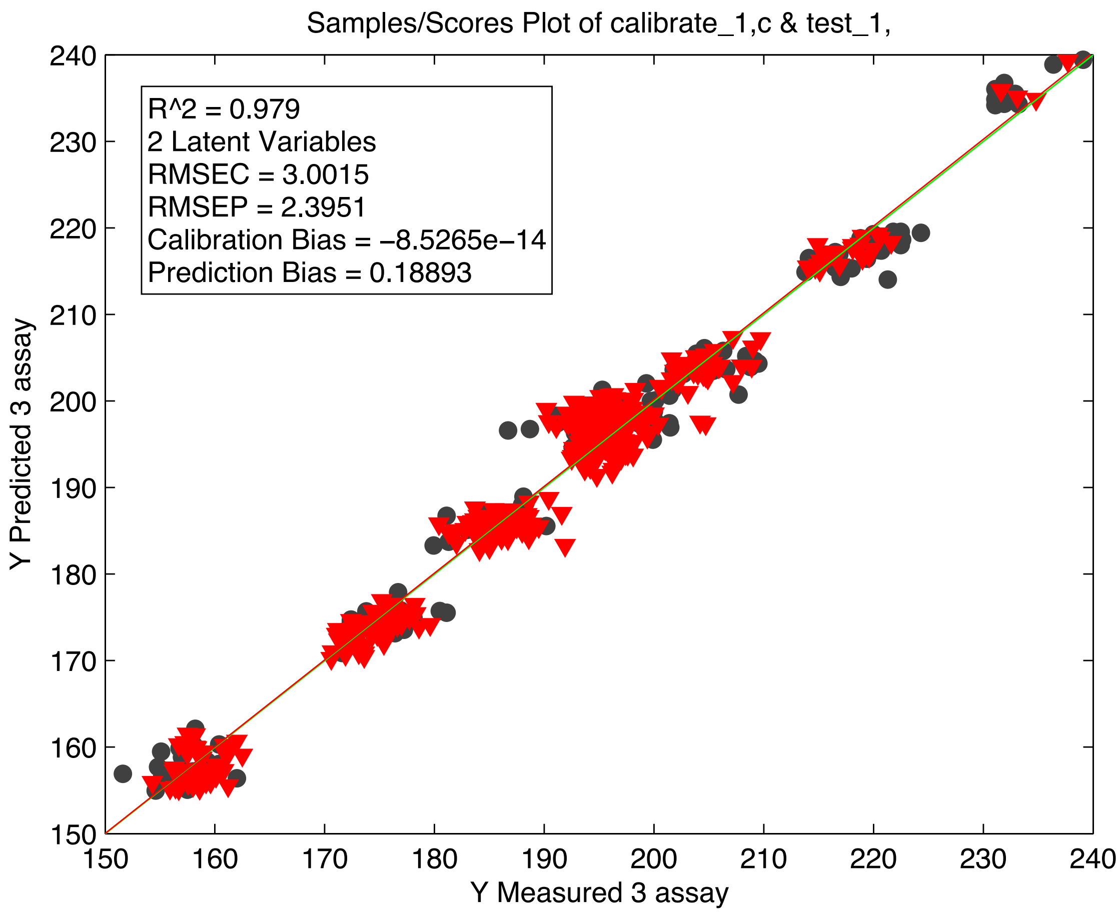


With MSC, GLS & MC



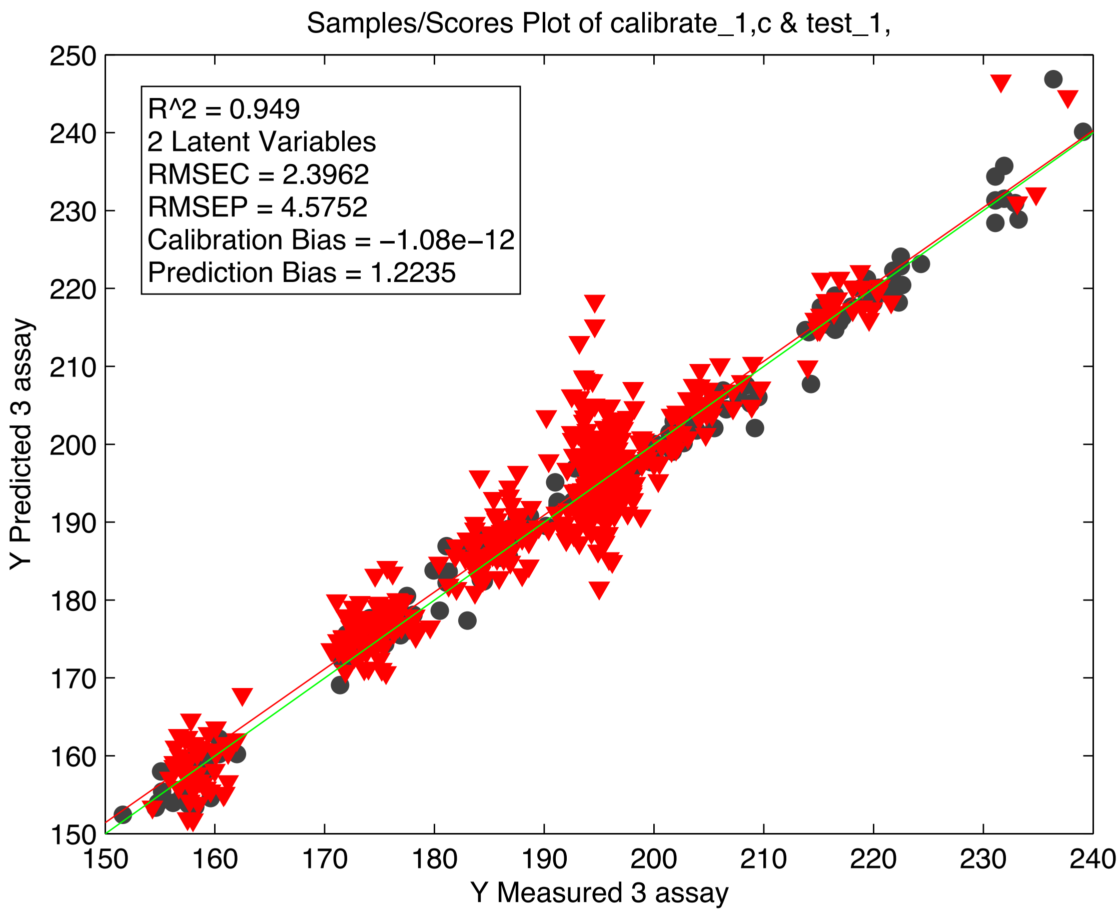
EIGENVECTOR RESEARCH INCORPORATED

With MSC, EPO & MC



EIGENVECTOR RESEARCH INCORPORATED

With MSC, ELS and MC



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 (mostly)

EIGENVECTOR RESEARCH INCORPORATED