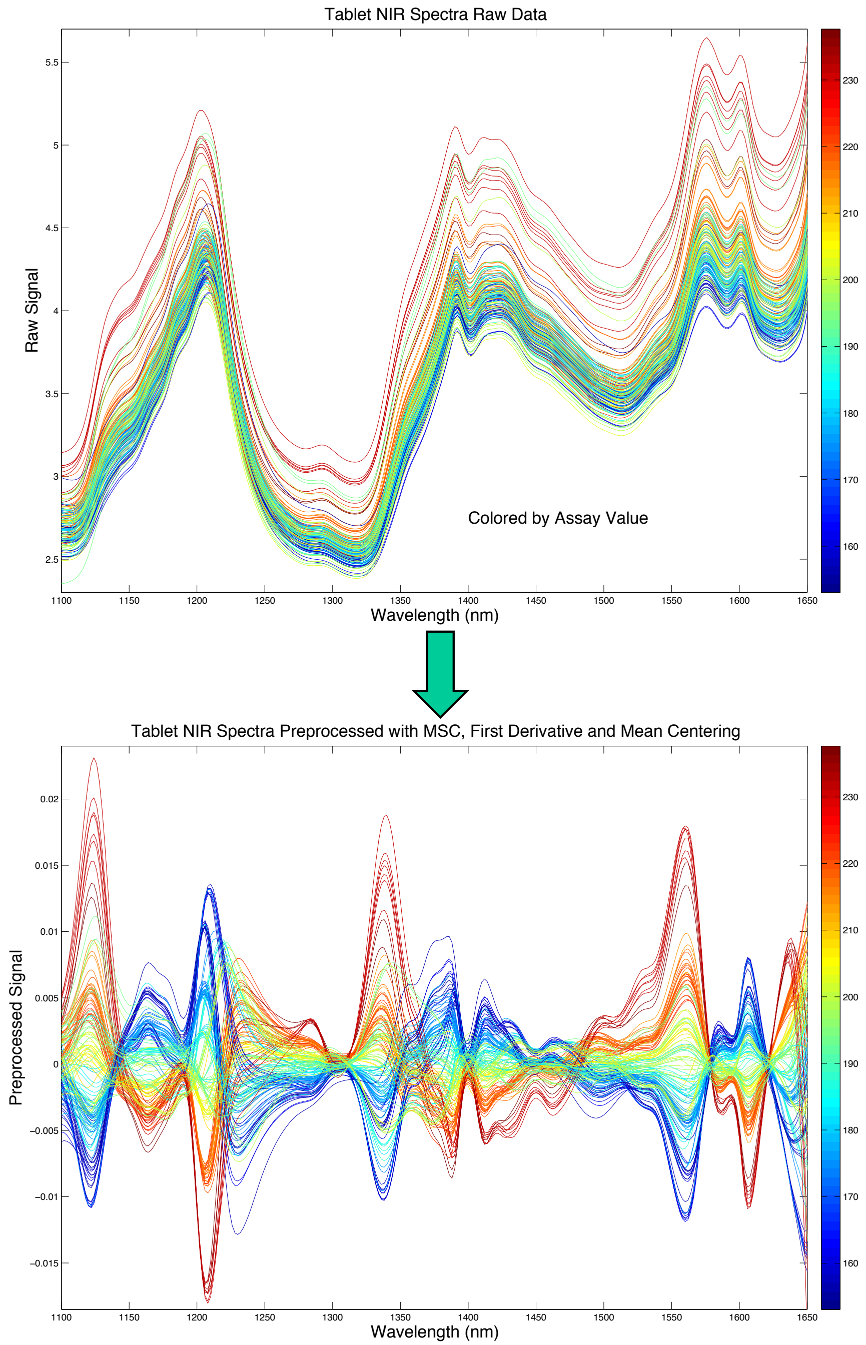
Data Preprocessing for Quantitative and Qualitative Models Based on NIR Spectroscopy

Barry M. Wise, Ph.D.

President
Eigenvector Research, Inc.
Manson, WA USA



EIGENVECTOR
RESEARCH INCORPORATED

Outline

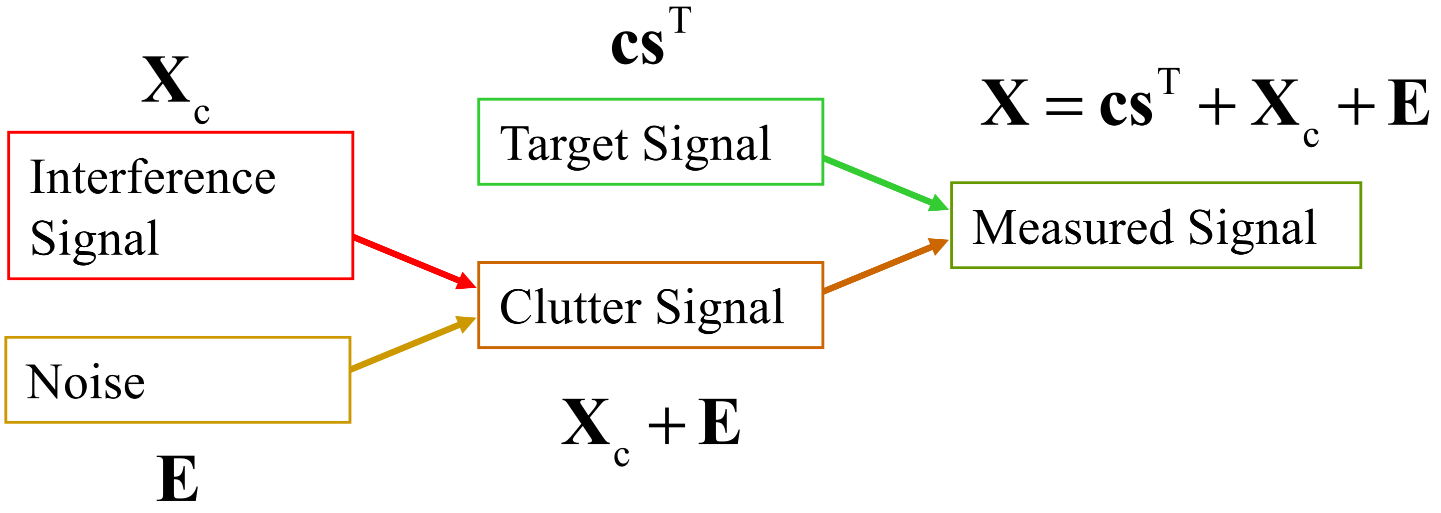
- Preprocessing Objective
- Definition of Clutter
- Linearization
- Mean Centering and Autoscaling
- Baseline Removal
- Normalization, Multiplicative Scatter Correction (MSC)
- Smoothing, Filtering and Derivatives
- Orthogonalization Filters: EPO, GLS
- Conclusions

Goal of Preprocessing

- Data preprocessing is what you do to the data *before* it hits the modeling algorithm (PCA, PLS, MCR, SIMCA, etc.)
- The goal of preprocessing is to remove variation you don't care about, *i.e. clutter*, in order to let the analysis focus on the variation you do care about
- Examples
  - Systems with scattering: physical vs. chemical effects
  - Classification: intra-class vs. inter-class variation

Measured Signal

- Clutter is present in all measurements (**X** & **Y**)
  - clutter = interferences + noise not of interest



$\mathbf{X}_c$

Interference Signal

Noise

$\mathbf{E}$

$\mathbf{cs}^T$

Target Signal

Clutter Signal

$\mathbf{X}_c + \mathbf{E}$

$$\mathbf{X} = \mathbf{cs}^T + \mathbf{X}_c + \mathbf{E}$$

Measured Signal

Sources of Clutter

- Systematic background variability
  - Variation in chemical interferents
  - Physical effects such as scattering due to particles
- Other changes in the system being observed
  - T, P changes, variable sample matrix, “dark current”
- Variance due to physics of instrument
  - e.g., drift, instrument changes, variable baseline or gain
  - Non-linearity, saturation
- Non-systematic random noise
  - homoscedastic, heteroscedastic

Reasons to Preprocess

- Reduces variance from extraneous sources
- Makes relevant variance more obvious
- Makes statistics work better
- Aids interpretation
- Avoids numerical problems

Transformation to Linear Form

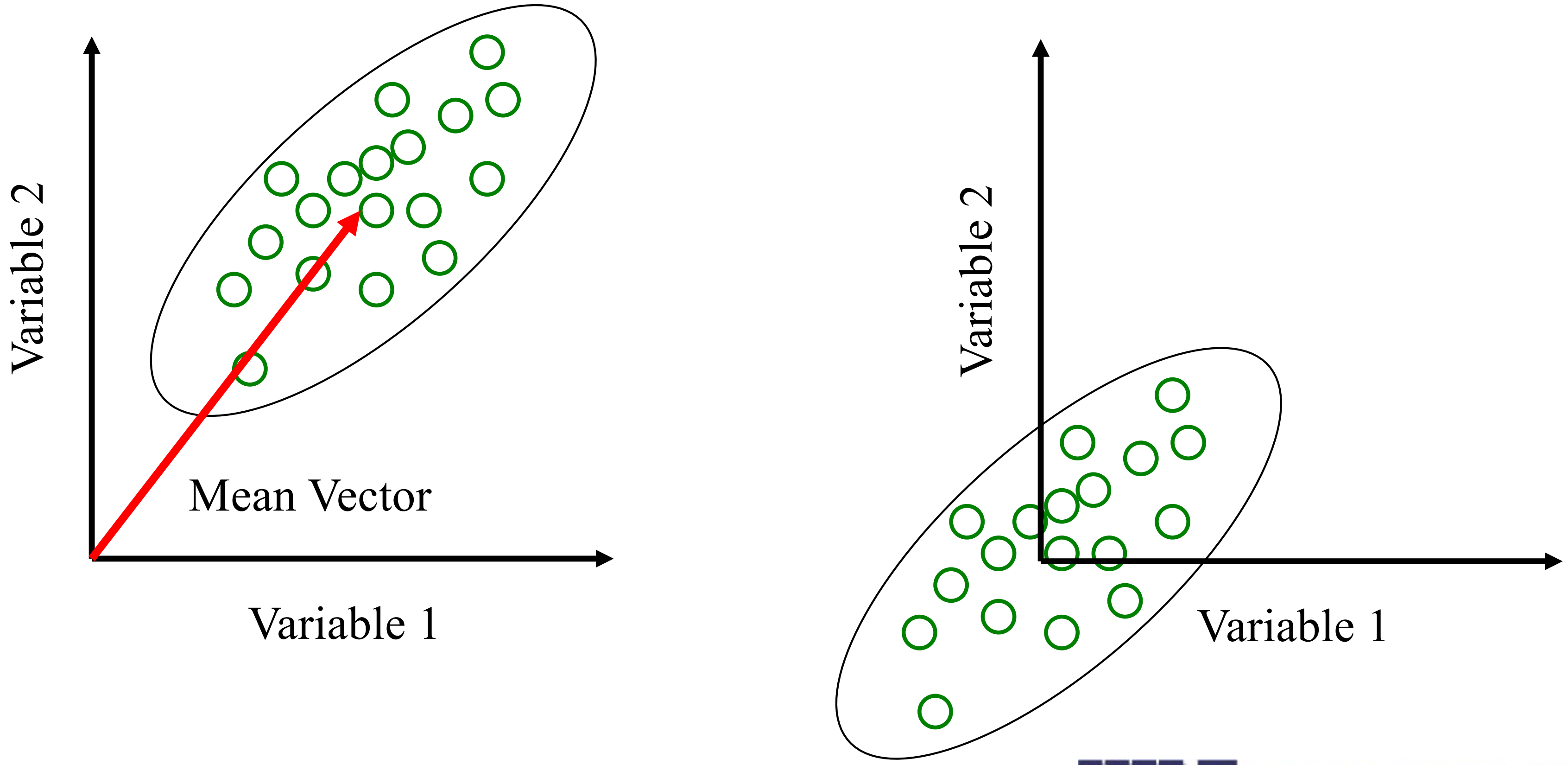
- Within X-block (predictor variables, *e.g.* spectra)
  - PCA works best with linear relationships
- Between X- & Y-block (predicted variable)
  - PLS regression assumes linear relation
- If possible, non-linear data should be converted to a linear form (e.g., use known physics of the system)
- Example:
  - Typically work with absorbance rather than transmittance
  - Log(I/I_0)

Mean Centering

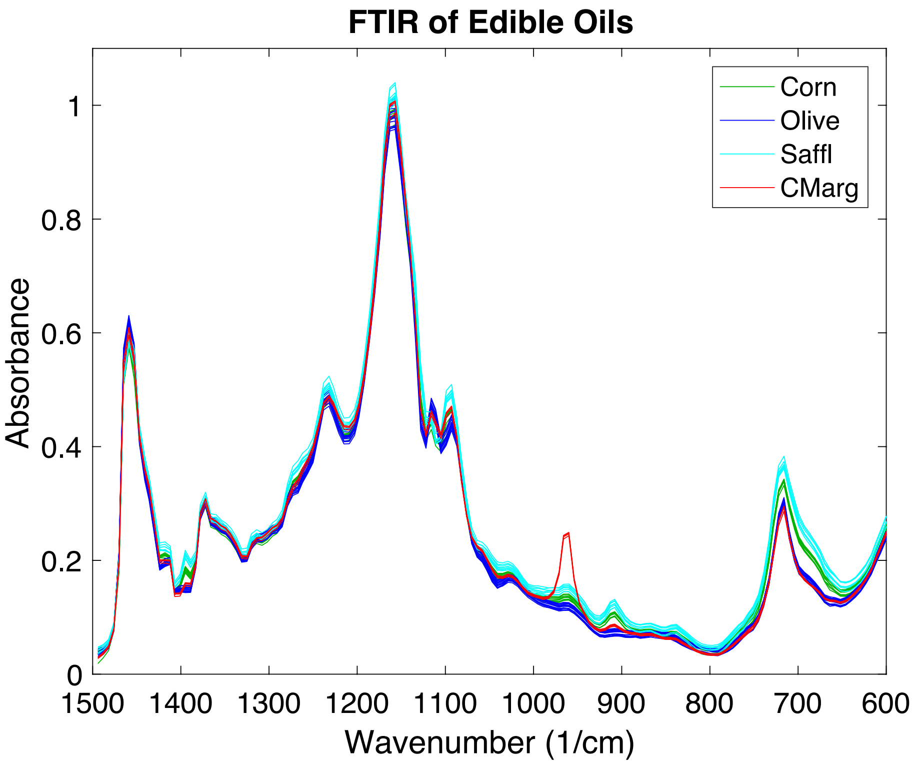
- Often we are most interested in how the data *varies* around the mean
- *Mean centering* is done by subtracting the mean off each column, thus forming a matrix where each column has mean of zero

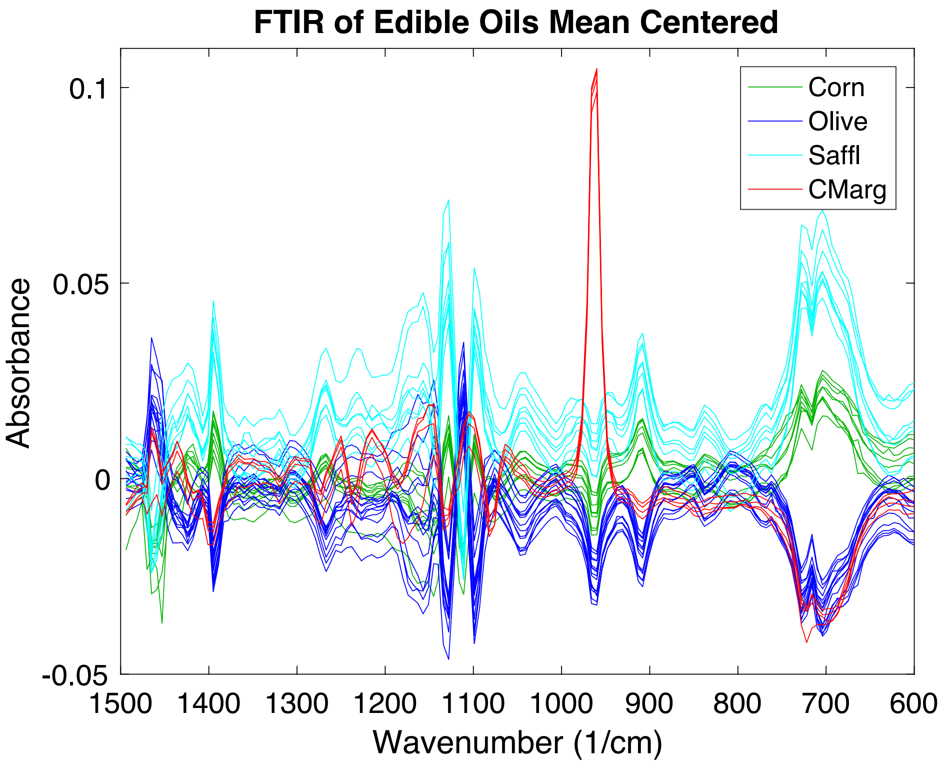
Centering is an Axis Translation

- Geometry for 2 variables



Mean Centering on Spectra

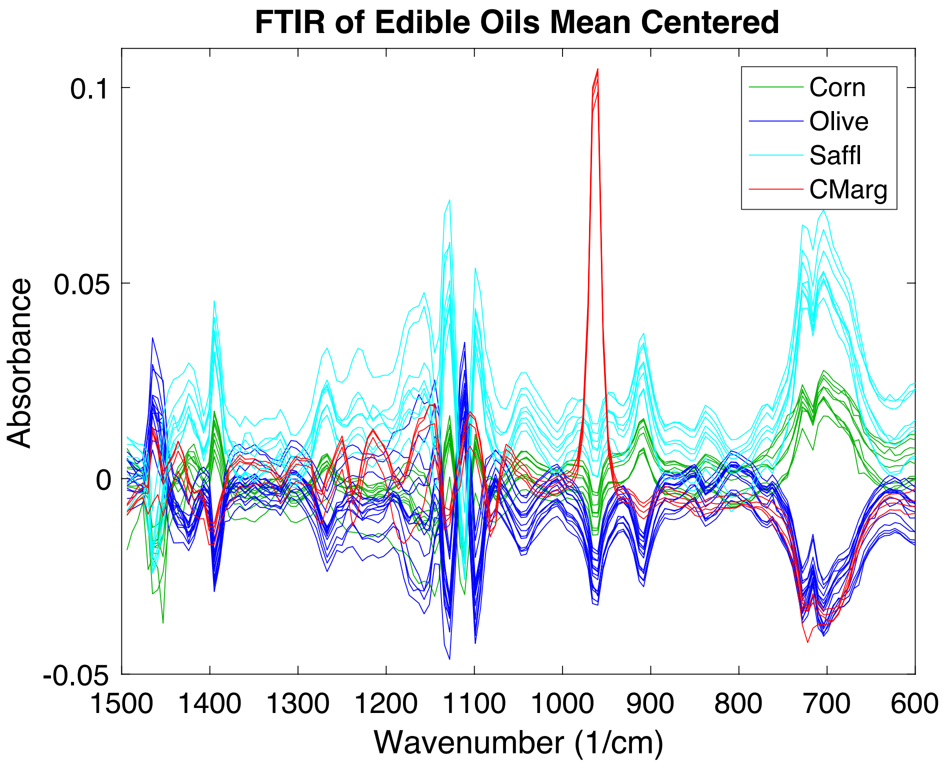


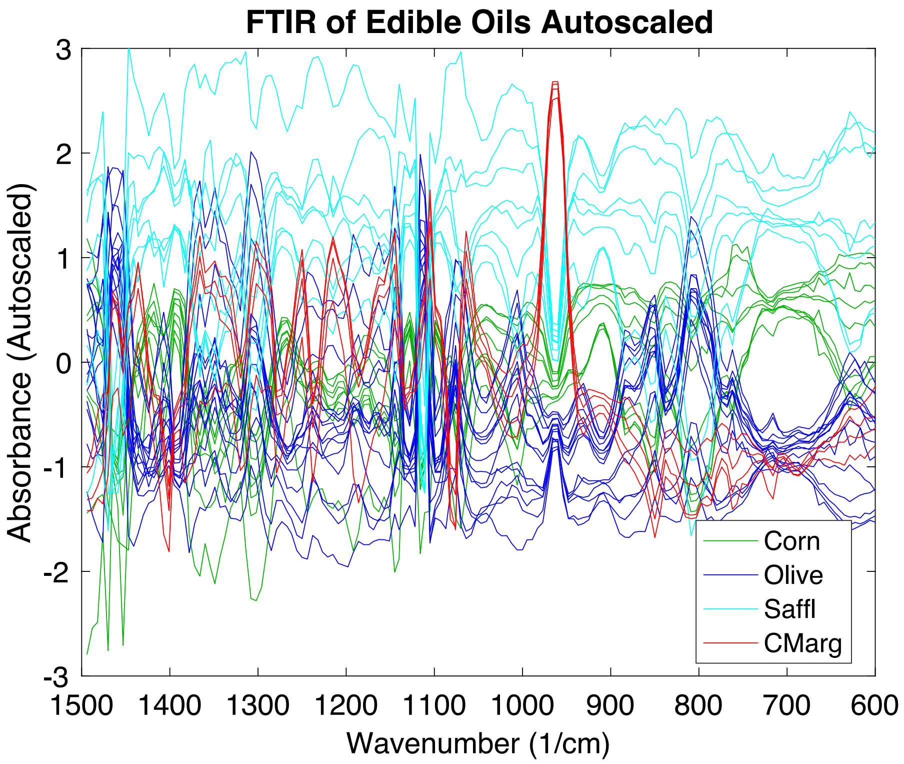


Variable Scaling

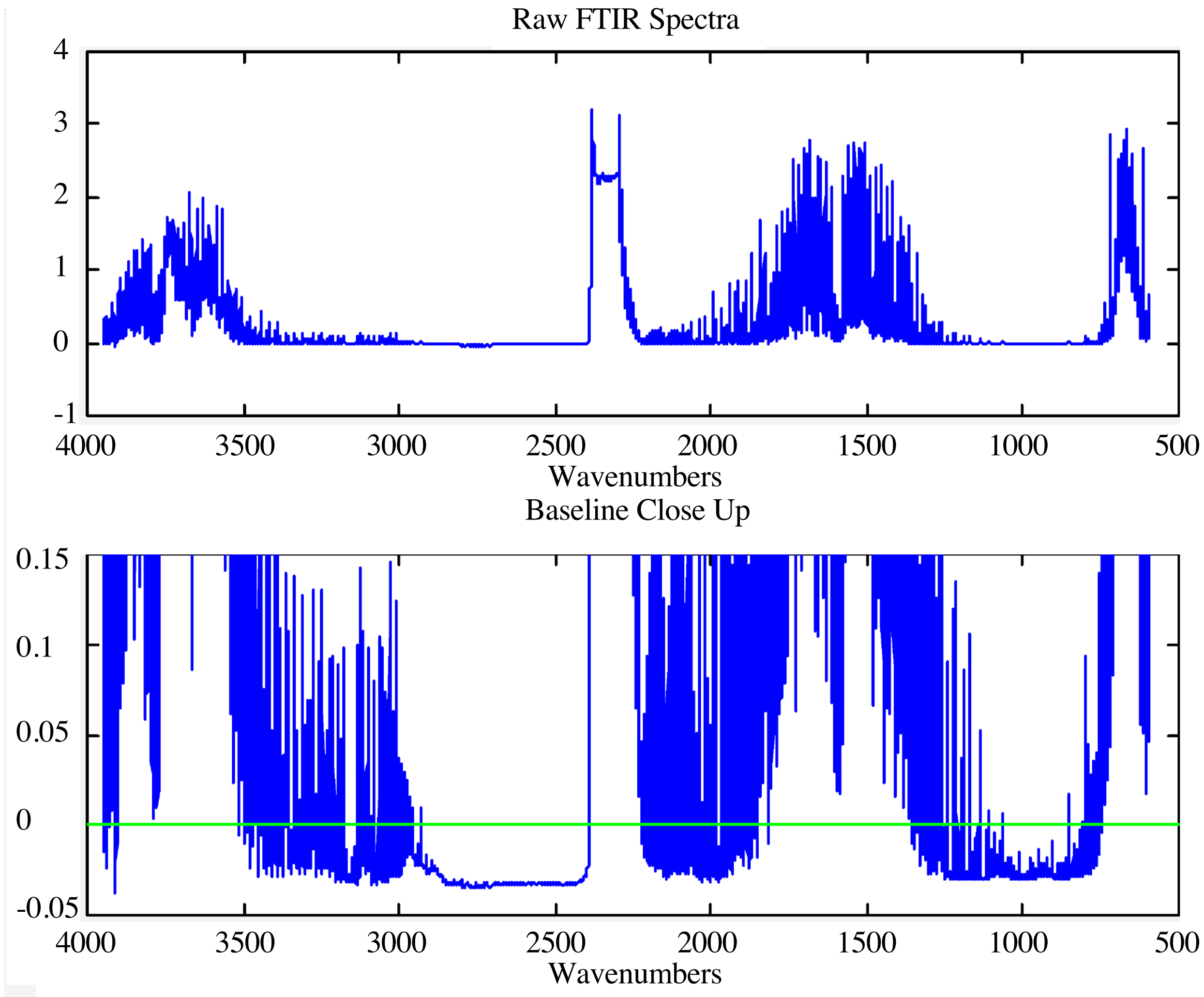
- Scaling is done to change variance of variables, and thus the weight given to them in modeling
- Most common is *autoscaling*, which makes variables unit variance and mean zero
  - Mean center variables
  - Divide by standard deviation
- Autoscaling removes all scale information
- What’s left is only how the variables correlate with each other
  - it is the “correlation matrix”

Autoscaling on Spectra


FTIR of Edible Oils Mean Centered


FTIR of Edible Oils Autoscaled

Sample-to-Sample Baseline



Baselines can exhibit simple offsets, slopes, polynomials or more complicated functions.

In the example, the offset is larger than the absorbance features of interest.

Adds variance that can inhibit predictive capability and make extraction of chemical information (e.g., via multivariate curve resolution) difficult.

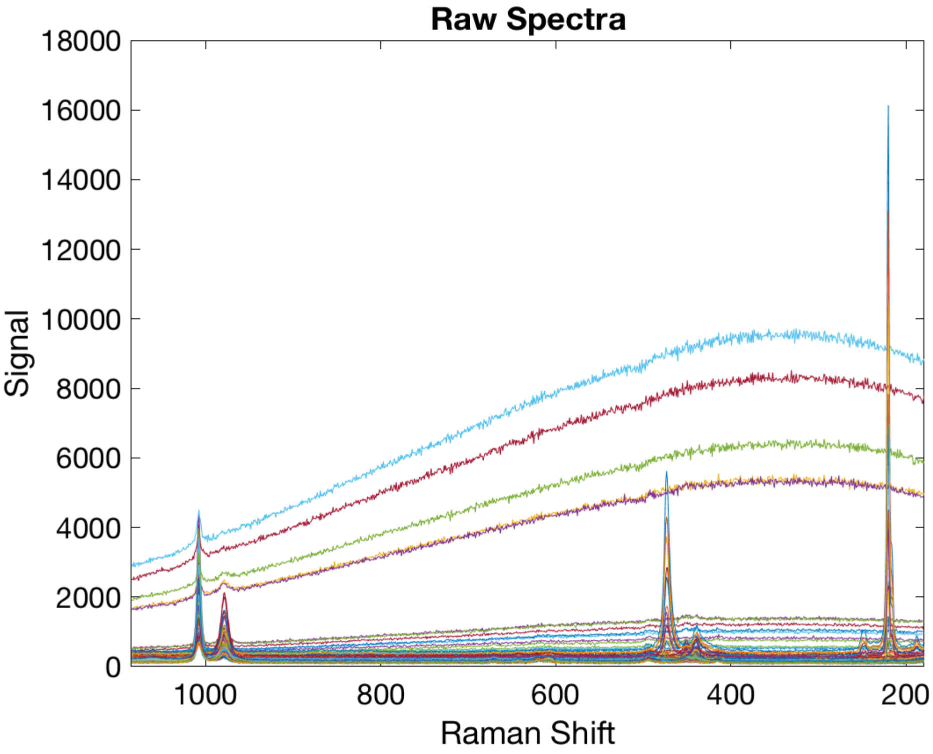
Background/Baseline Subtraction

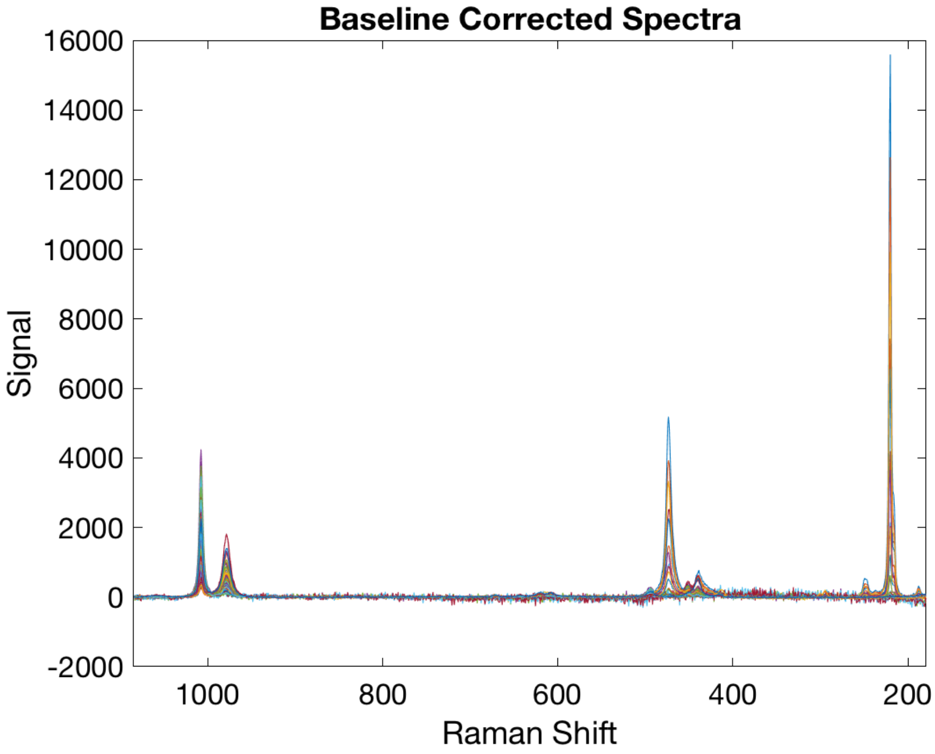
Removal of broad (low-frequency) interferences while retaining higher-frequency features. Only low-order polynomials are used to model the background.

- **Detrend:** fit polynomial to *entire* spectrum
- **Selected-Points baselining:** fit polynomial to selected points in spectrum
- **Weighted Least-squares (a.k.a. asymmetric) baselining:** fit to *automatically* selected points on the bottom of the spectrum
- **Windowed:** Whittaker, Rolling Ball, Median, Minimum, etc.
- **Etc.**

Selected-Points Baseline

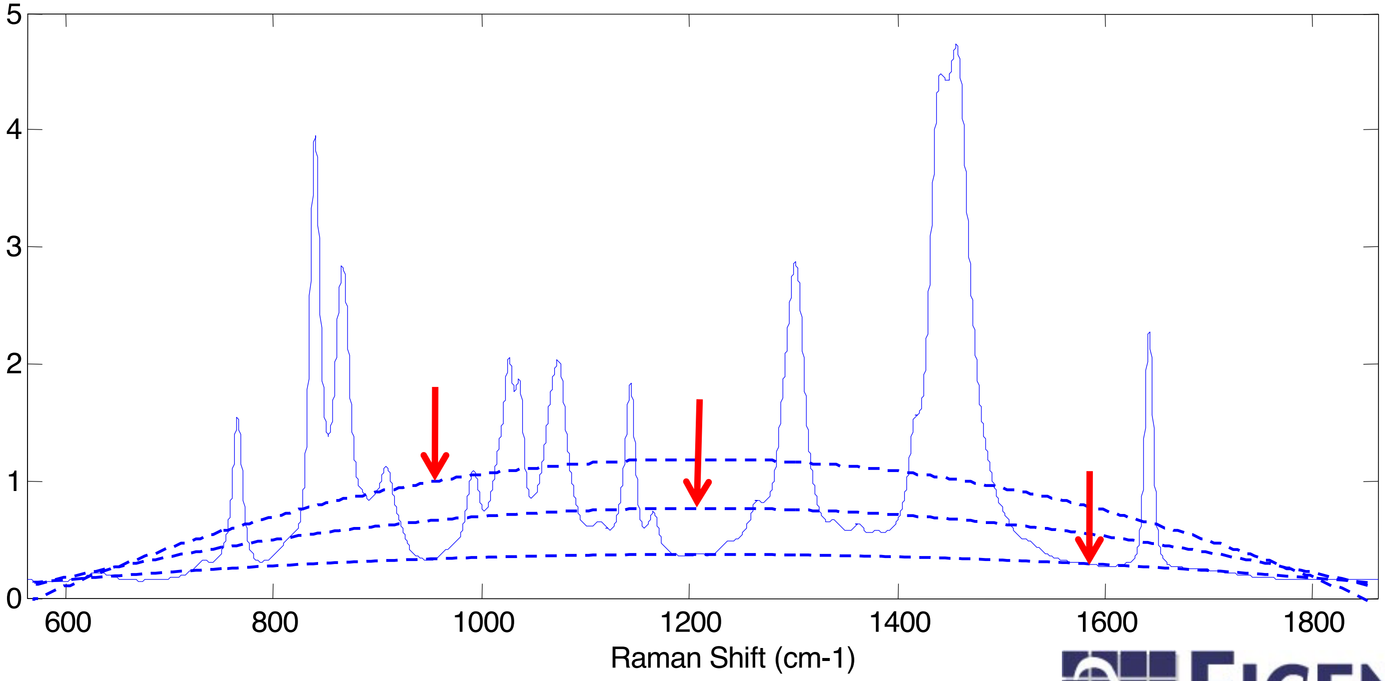
- Detrend based on points in spectrum known to be only baseline. Subtract the result from all channels.
  - good when zero points are known a priori





Weighted Least-Squares Baselining

- Automatic selection of baseline points by fitting polynomial to the “bottom” (or “top”) of the spectrum → asymmetric fit.
  - Starts with a fit to all points (detrend) then de-weights points above the baseline (those with large positive residuals).
  - Iterates until only points w/in a defined tolerance on the residuals are kept. (Need to define tolerance on the residuals.)
  - Easy approach for simple baselines (e.g., polynomials).
  - Can also include known baseline functions.



EIGENVECTOR RESEARCH INCORPORATED

Sample Normalization Methods

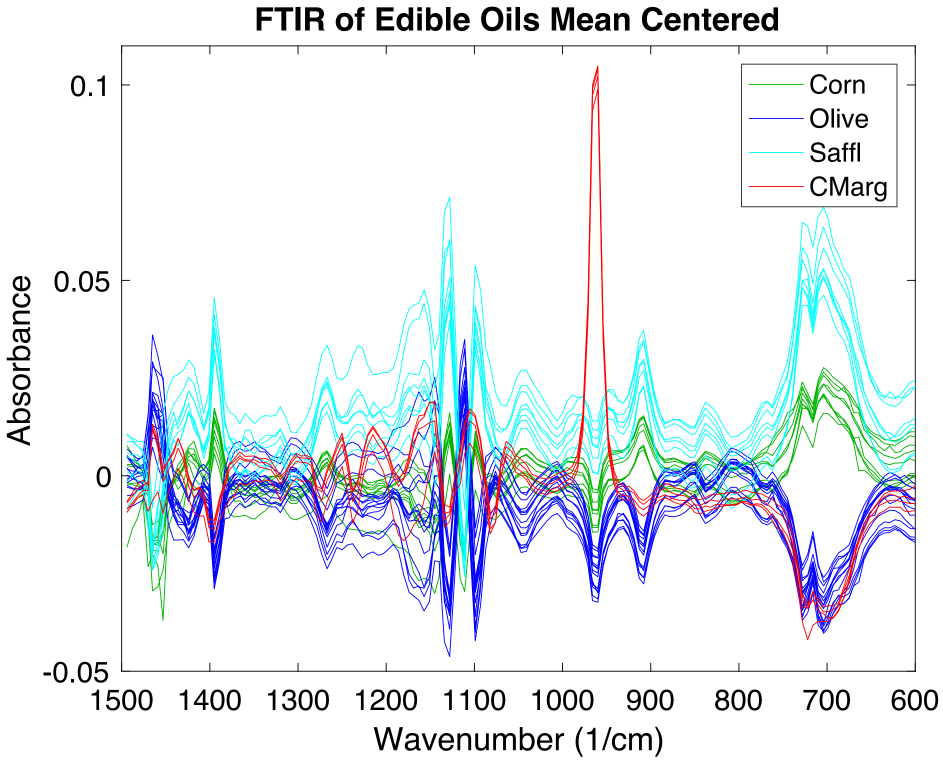
- Previous examples removed an offset. How is variance due to changing magnitude removed?
  - variable source or lighting magnitude
  - scattering effects
- **Row Normalization**: removes magnitude
- **Standard Normal Variate (SNV)**: subtracts the row mean from each row and scales to unit variance
  - Autoscaling of the rows
- **Multiplicative Scatter Correction:** Determines scale factor that best fits new spectrum to reference
- Be aware that these can “blow up” low signal noisy samples to have more variance

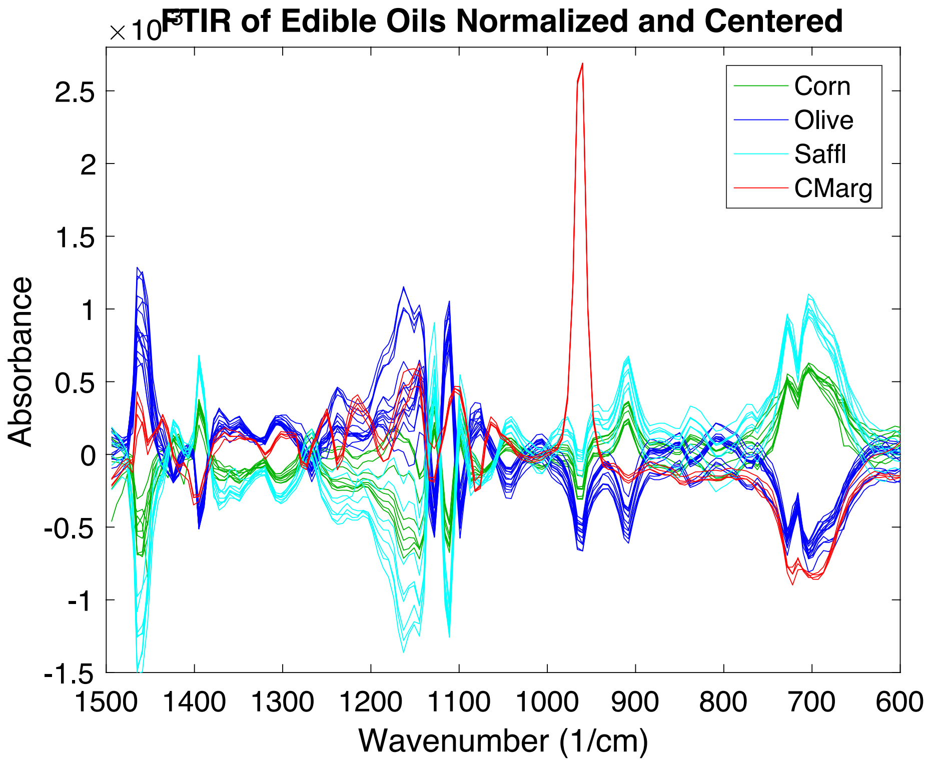
Normalization

- Normalize each row / spectrum
- Order of normalization (*p*-norm)
  - 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

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

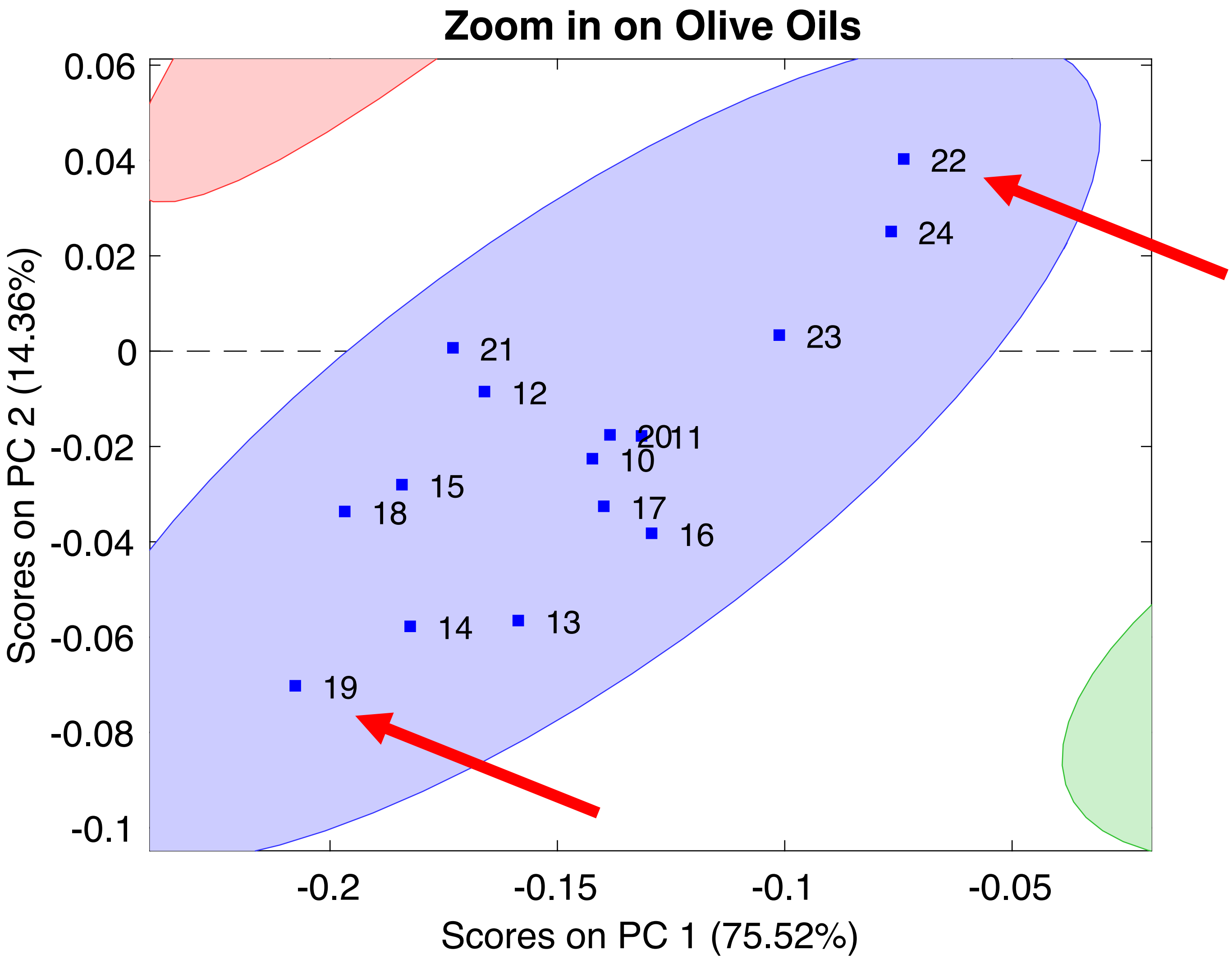

FTIR of Edible Oils Mean Centered


FTIR of Edible Oils Normalized and Centered

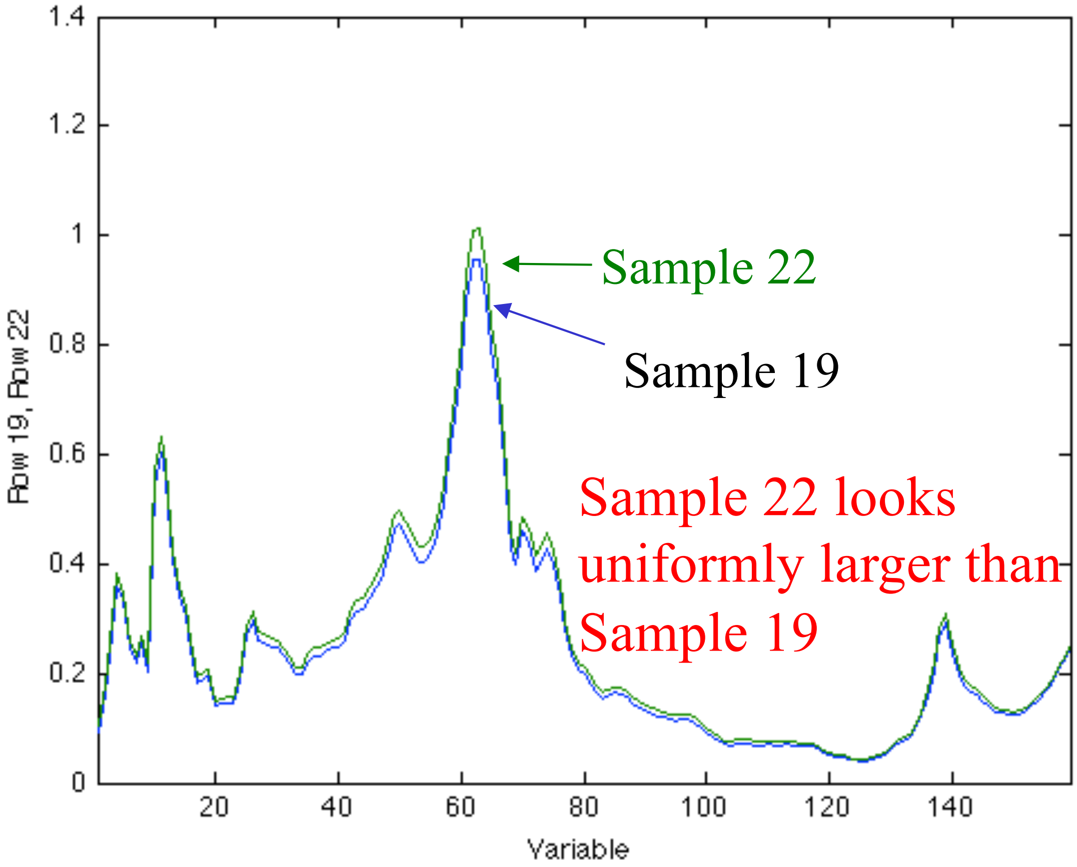
Scatter / Signal Correction

- Multiplicative Scatter Correction (MSC)
  - Attempts to remove offset *and* row magnitude variability
  - Result is less signal related to scattering artifacts and more signal related to analyte(s) of interest

MSC Example


Zoom in on Olive Oils
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
Scores on PC 2 (14.36%)
-0.2
-0.15
-0.1
-0.05
Scores on PC 1 (75.52%)
22
24
23
21
12
2011
10
15
18
17
16
14
13
19

Spectra (Selected Wavelengths)
Samples 19 & 22

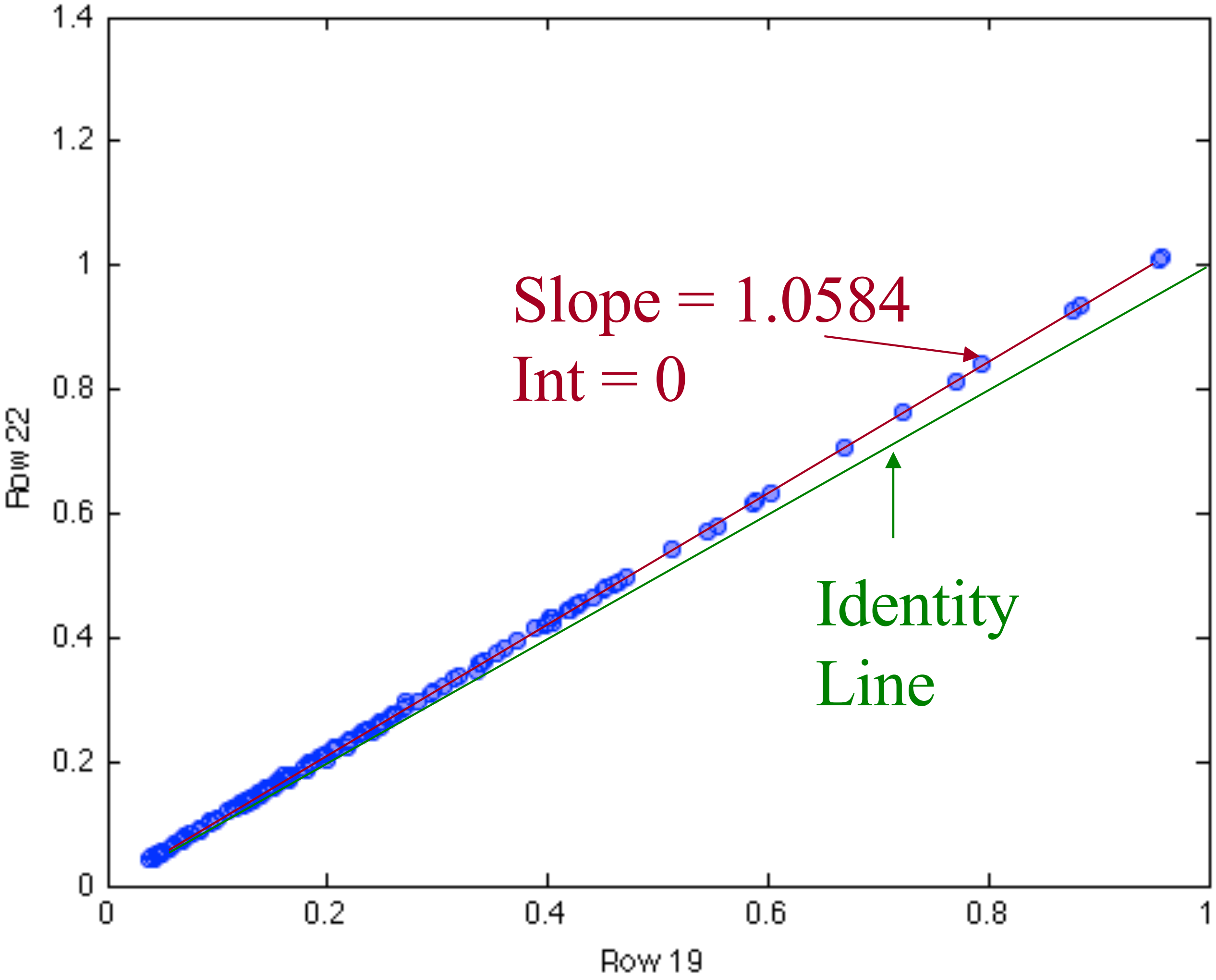


Sample 22

Sample 19

Sample 22 looks uniformly larger than Sample 19

Plot Sample 22 vs. Sample 19



Slope = 1.0584
Int = 0

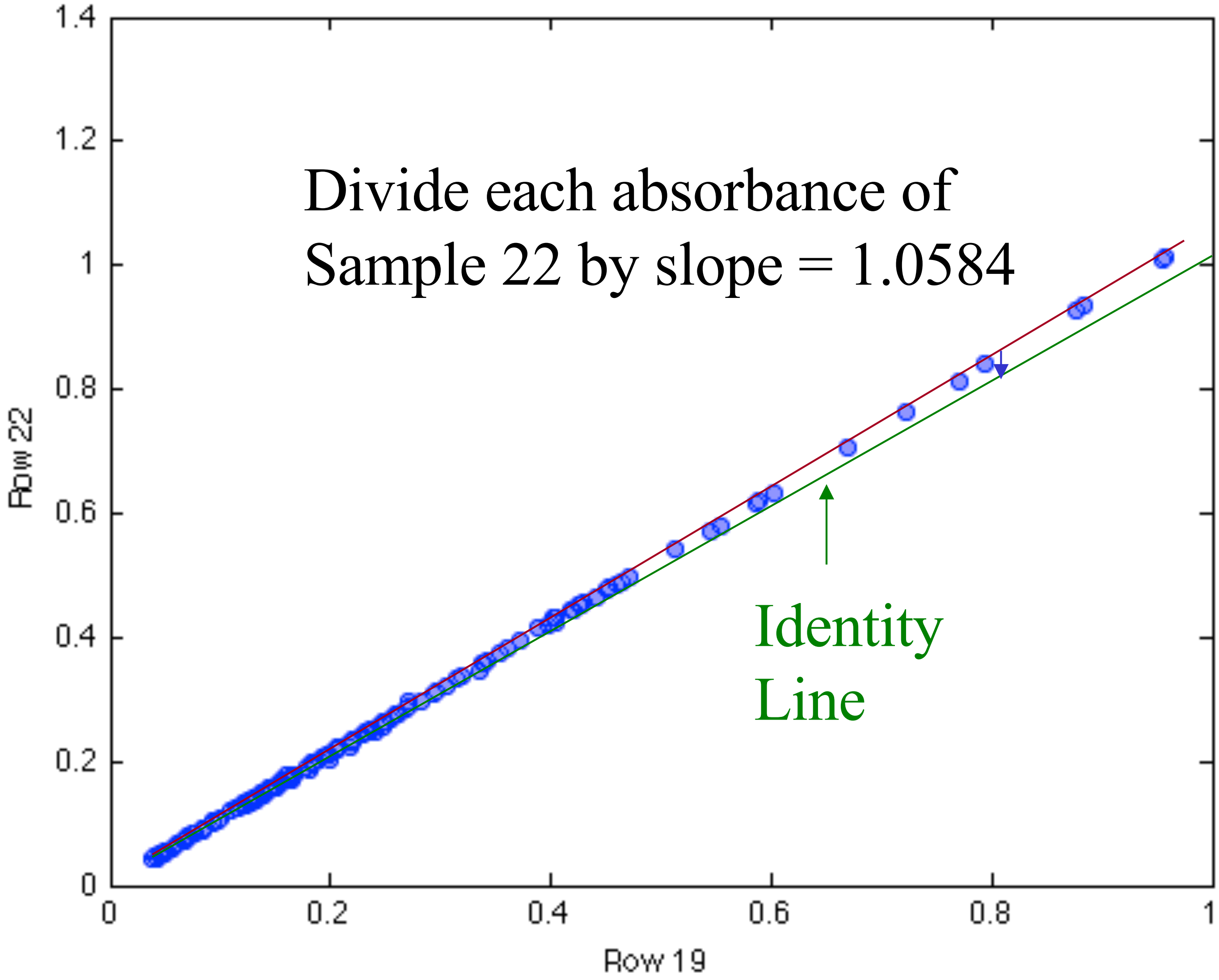
Identity Line

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

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

MSC

Multiplicative Signal (Scatter) Correction

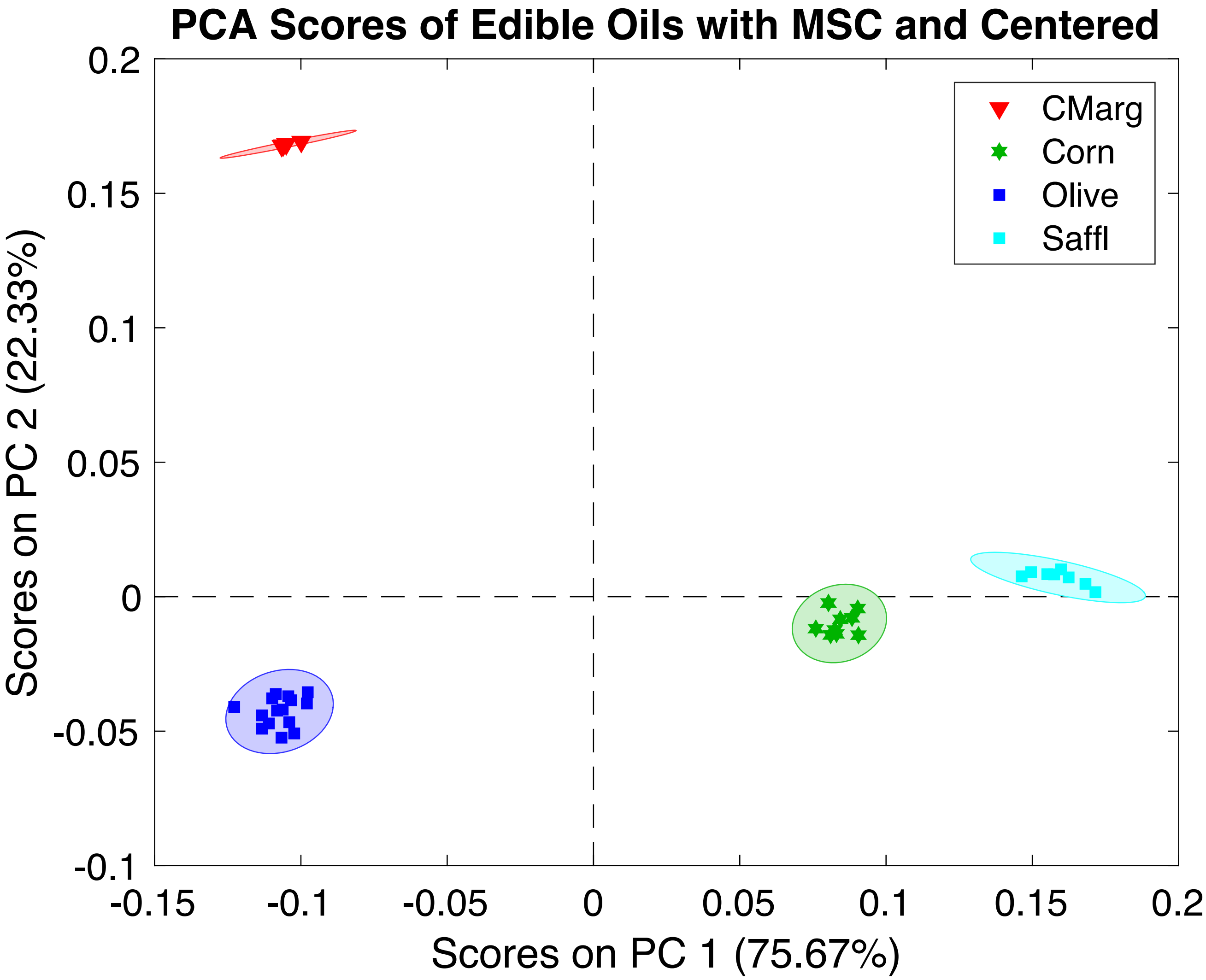


Divide each absorbance of Sample 22 by slope = 1.0584

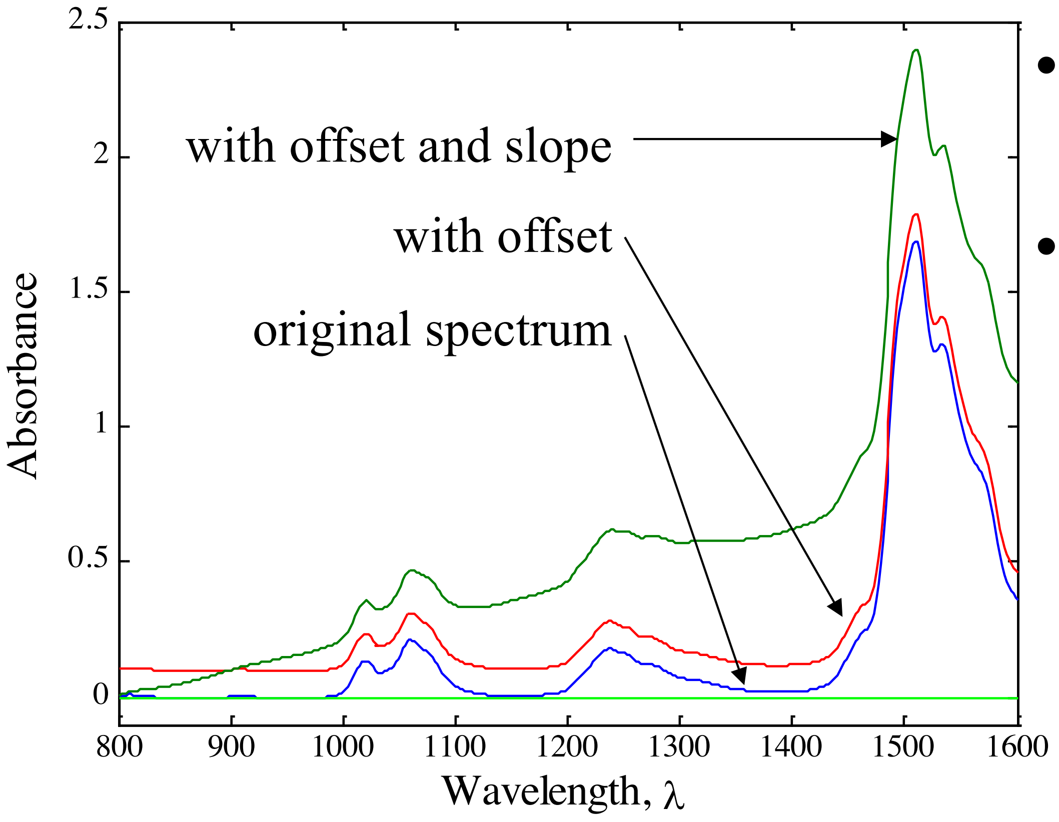
Identity Line

EIGENVECTOR RESEARCH INCORPORATED

With MSC

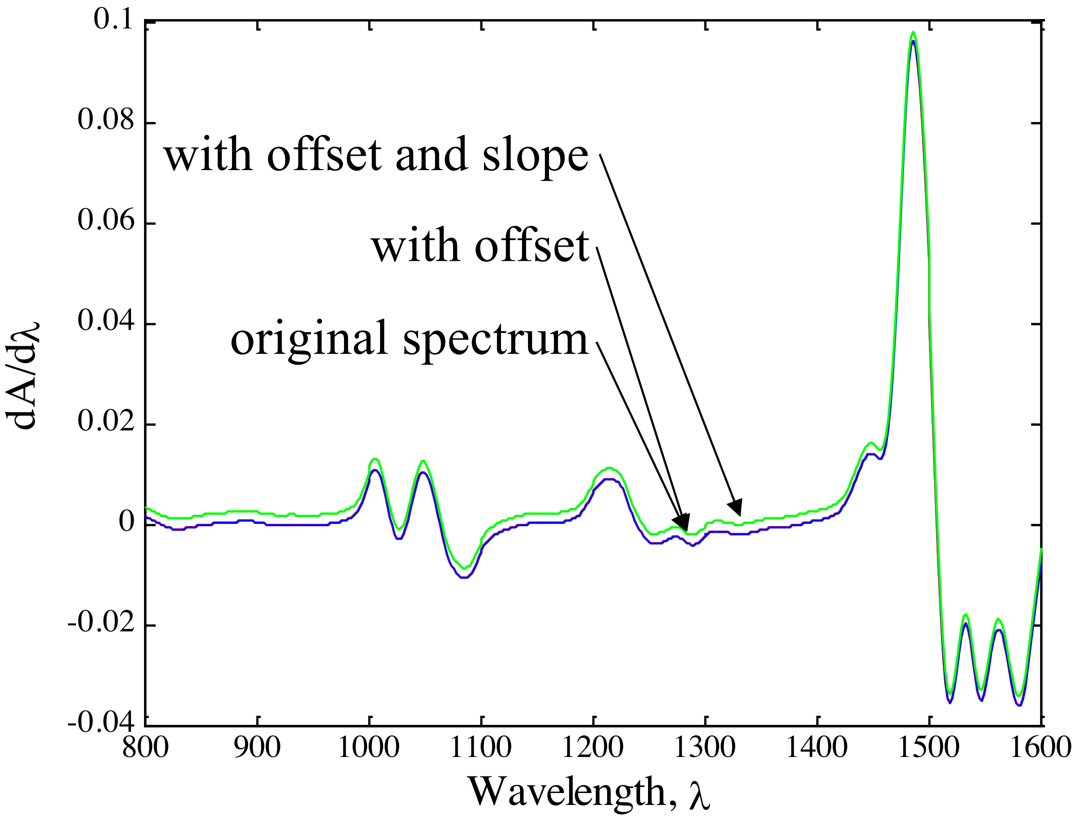

PCA Scores of Edible Oils with MSC and Centered
CMarg
Corn
Olive
Saffl
Scores on PC 2 (22.33%)
Scores on PC 1 (75.67%)
0.2
0.15
0.1
0.05
0
-0.05
-0.1
-0.15
-0.1
-0.05
0
0.05
0.1
0.15
0.2

Savitzky-Golay Smoothing and Derivatives



- Derivatives wrt λ can be used to remove offsets/slopes
- Savitzky-Golay smoothing and derivatives
  - piece-wise fit of polynomials to each spectrum
  - use fit directly for smoothing
  - use derivative in each window for estimate of derivative wrt λ
  - smooth + derivative can be boiled down to a set of coefficients

Savitzky-Golay First Derivative



multicomponent Beer’s Law

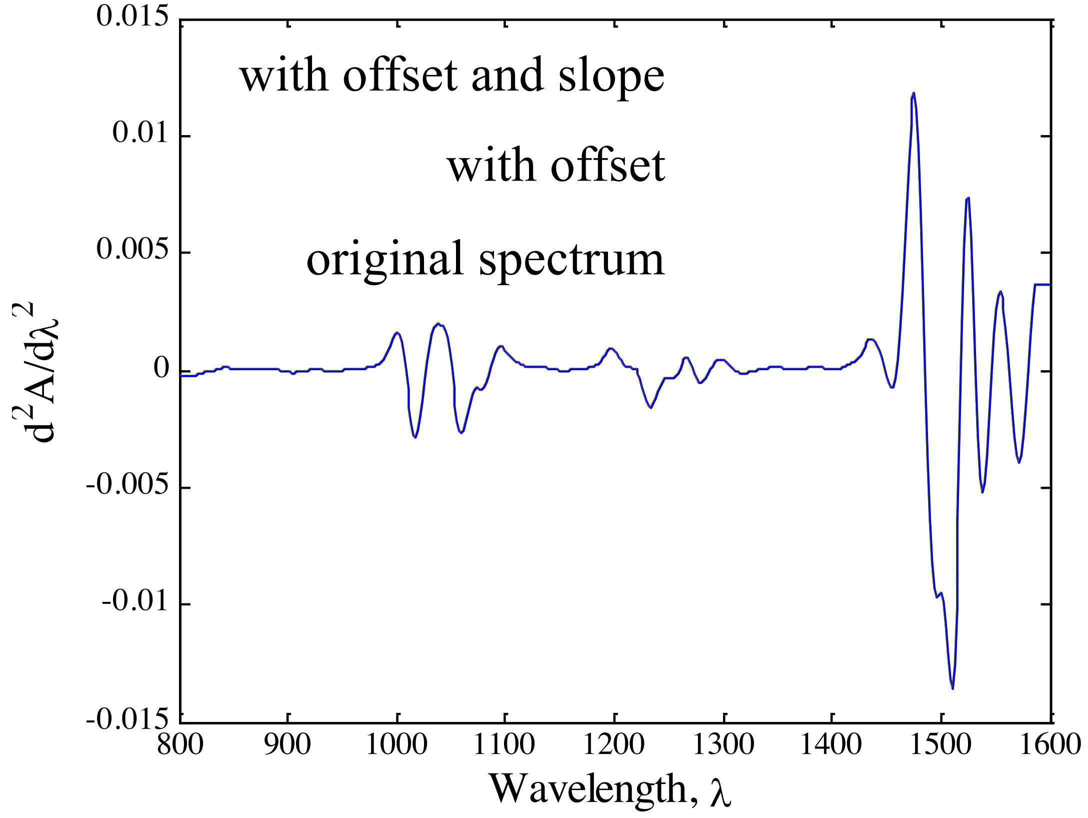
$$\mathbf{x} = \mathbf{c}\mathbf{S}^{\mathrm{T}}$$

first derivative removes the offset

$$\mathbf{x} = \mathbf{c}\mathbf{S}^{\mathrm{T}} + \alpha\mathbf{1}^{\mathrm{T}}$$

$$\frac{d\mathbf{x}}{d\lambda} = \mathbf{c}\frac{d\mathbf{S}^{\mathrm{T}}}{d\lambda}$$

Savitzky-Golay Second Derivative



multicomponent Beer’s Law

$$\mathbf{x} = \mathbf{c}\mathbf{S}^{\mathrm{T}}$$

second derivative remove the offset and slope

$$\mathbf{x} = \mathbf{c}\mathbf{S}^{\mathrm{T}} + \alpha\mathbf{1}^{\mathrm{T}} + \beta\boldsymbol{\lambda}$$

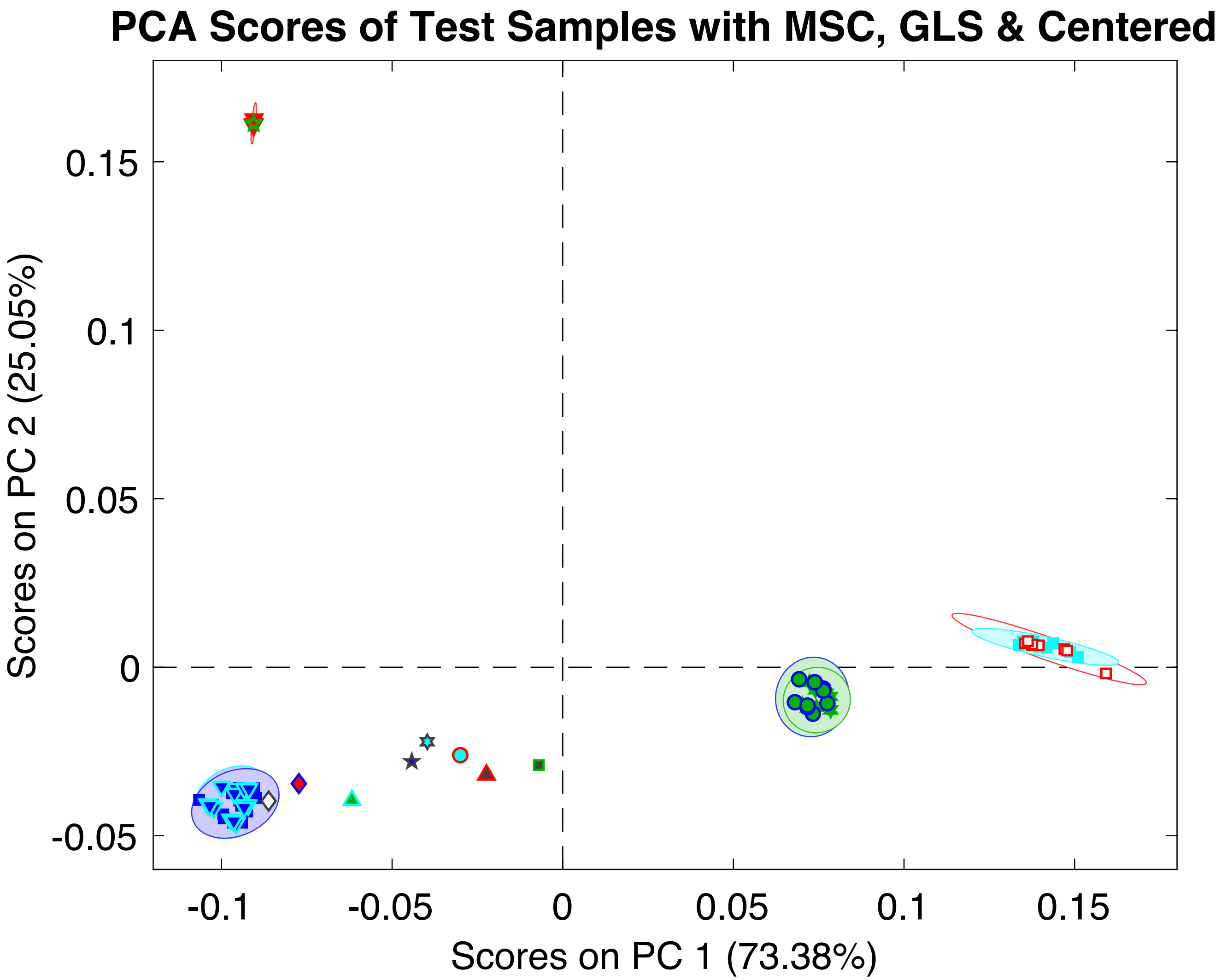
$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}\lambda} = \mathbf{c}\frac{\mathrm{d}\mathbf{S}^{\mathrm{T}}}{\mathrm{d}\lambda} + \beta$$

$$\frac{\mathrm{d}^2\mathbf{x}}{\mathrm{d}\lambda^2} = \mathbf{c}\frac{\mathrm{d}^2\mathbf{S}^{\mathrm{T}}}{\mathrm{d}\lambda^2}$$

EPO and GLS Filters

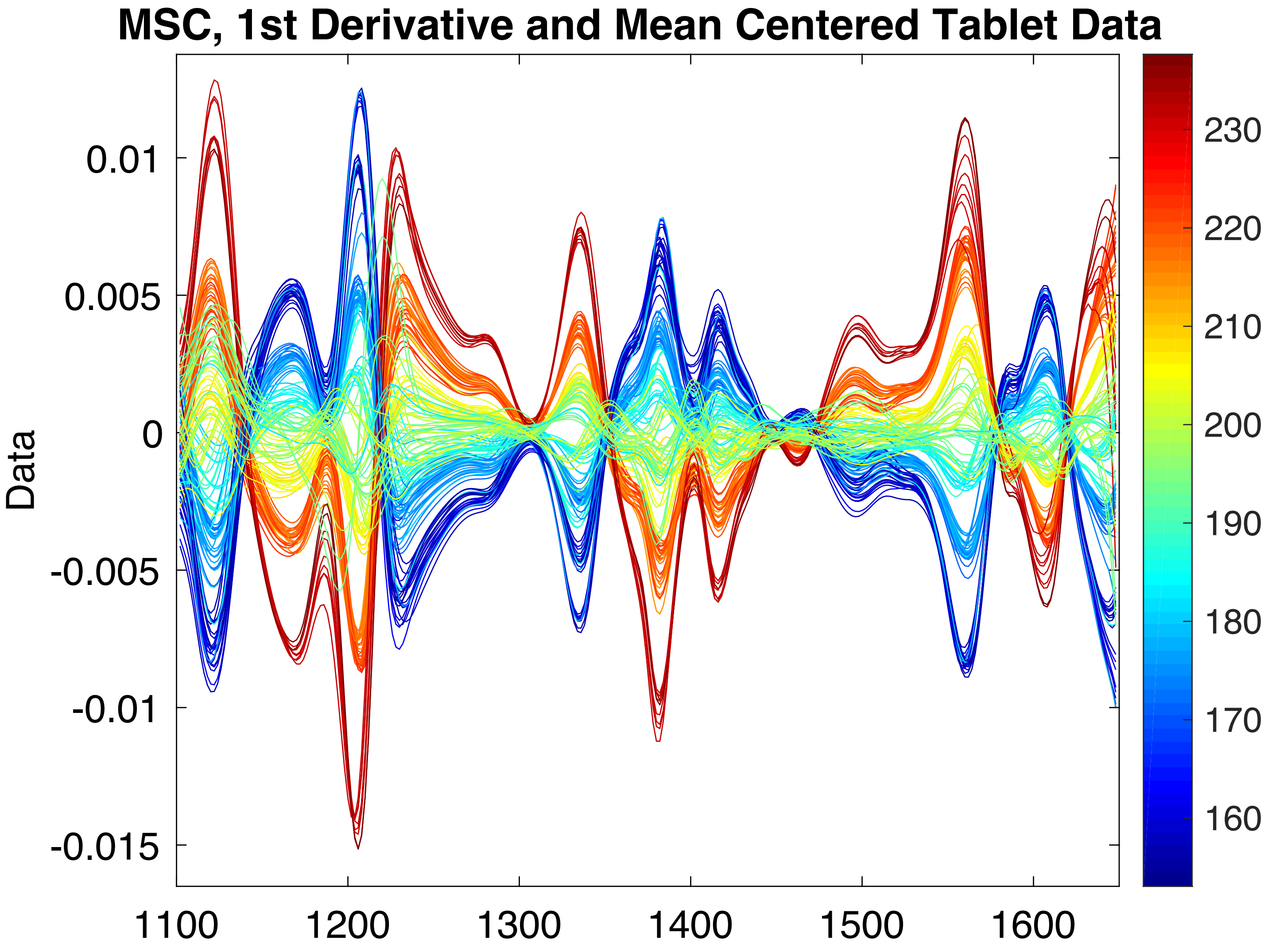
- EPO = External Parameter Orthogonalization
- GLS = Generalized Least Squares filter
- Both use samples that characterize the clutter
  - Variation not related to the problem of interest
  - Classification problems: inter-class variance
  - Regression problems: samples with same property
- EPO makes PCA model of clutter, orthogonalizes data against first few PCs – hard filter
- GLS calculates weighted inverse of clutter covariance, applies to all data – soft filter

With MSC and GLS

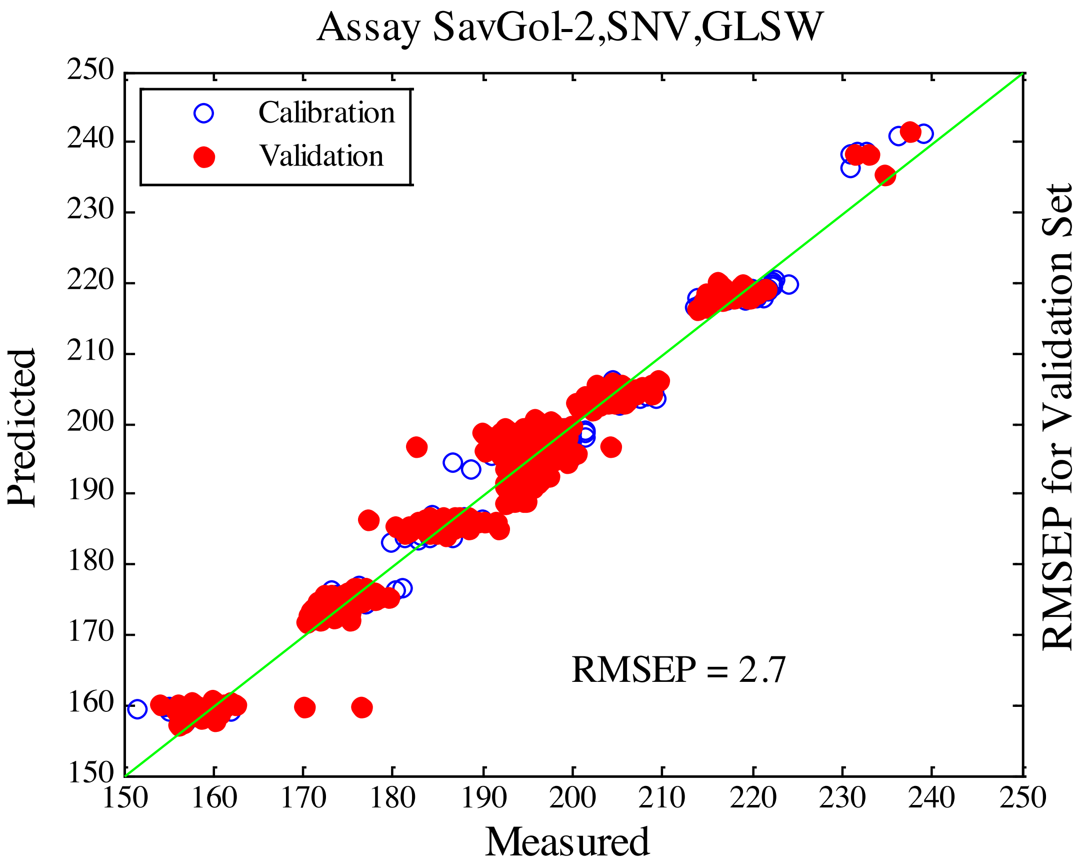


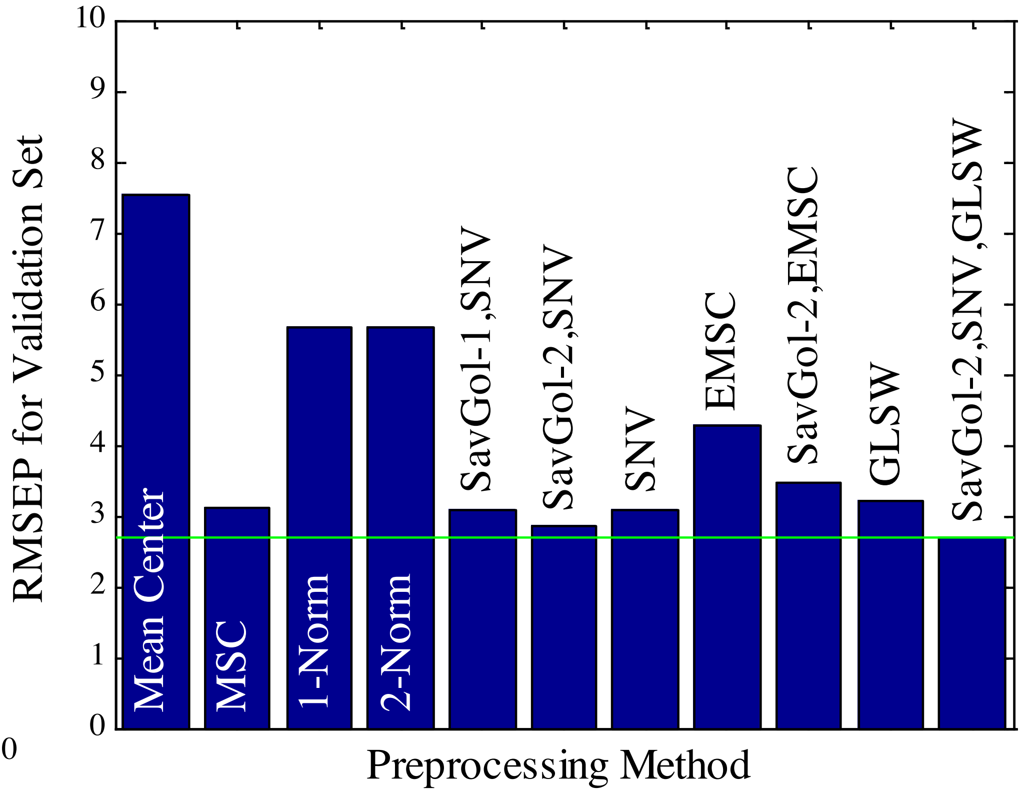
NIR Shootout 2002

- Estimate tablet assay value from NIR transmittance
  - Calibration (155 samples), Test (460 samples)



Prediction Error on Validation Set





EIGENVECTOR
RESEARCH INCORPORATED

Perspectives on Preprocessing

- Order matters. The general approach is:
  1. Background and offset removal
  2. Normalization
  3. Centering
  4. Scaling
- Always keep in mind: "what is each preprocessing step supposed to be doing?...."
- Plot data after pre-preprocessing and color code!
- Always compare the effect of the pre-processing (classification or regression error rates) with the results from a model based on the raw data

Pre-processing will offer…

- Models with better predictive or classification performance and/or
- Simpler models that are more robust and/or more easy to interpret
- But there is a risk that you can remove useful information from data
- Preprocessing must be validated as part of the model development process