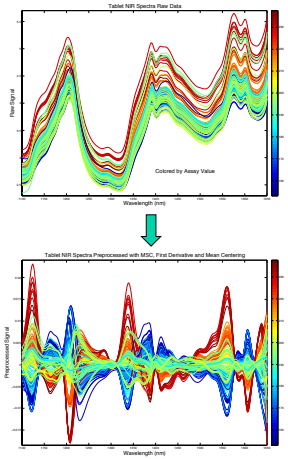
Advanced Preprocessing



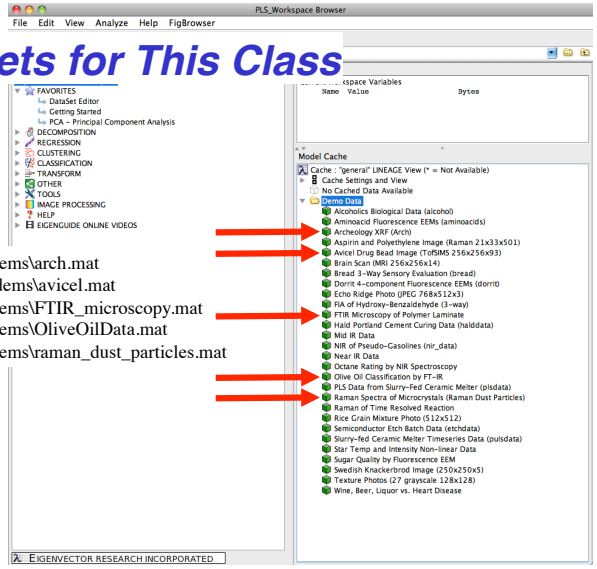
EIGENVECTOR RESEARCH INCORPORATED

Outline

- Software and Data Sets
- Preprocessing Objective
- Motivation: Simple Example with Mean Centering ???
- Review: Mean Centering and Autoscaling
- Baseline Removal
- Standard Normal Variate, Normalization, Scatter Correction (MSC)
- Smoothing and Filtering, Savitzky-Golay
- Derivatives
- Scaling: Autoscaling with Offset, Poisson, Exponential Decay
- Orthogonalization Filters: OSC, O-PLS, GLS
- Linearizing, Matrix Rank and the Bilinear Model
- The Extended Mixture Model (ELS)
  - Target Detection, Classical Least Squares, MSC and Extended Multiplicative Scatter Correction, Extended Least Squares in Curve Resolution, Generalized Least Squares
- Scaling for Multi-block data
- Preprocessing order

EIGENVECTOR RESEARCH INCORPORATED

Data Sets for This Class



… \PLS_Toolbox\dems\arch.mat
… \MIA_Toolbox\dems\avicel.mat
… \PLS_Toolbox\dems\FTIR_microscopy.mat
… \PLS_Toolbox\dems\OliveOilData.mat
… \PLS_Toolbox\dems\raman_dust_particles.mat

Why Preprocess?

- What's the objective of the analysis?
  - Simple example: Compare new measurements to a model of a system or process
  - The system might be characterized as the mean and variance about the mean
    - multivariate statistical process control
    - anomaly detection
  - In this case, the data are centered to the mean of
    - normal process data
    - null / non-anomaly data

EIGENVECTOR RESEARCH INCORPORATED

Mean-Centering

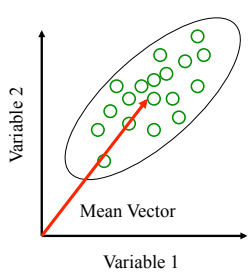
- Detection was relative to the mean …
- Mean-centering is used to
  - allow PCA models to capture variance about the mean
    - exploratory analysis
    - MSPC (assumed stationary, as in the detection models)
    - SIMCA (classification based on distance from cluster mean)
    - other models …
  - calibration (normal, null) data are centered to it's mean and new (test) data are centered to that mean
    - assumes stationary process
  - avoid numerical problems

Seasholtz, M.B., and Kowalski, B.R., "The Effect of Mean Centering on Prediction in Multivariate Calibration", *J. Chemometr.*, **6**, 103-111 (1992).

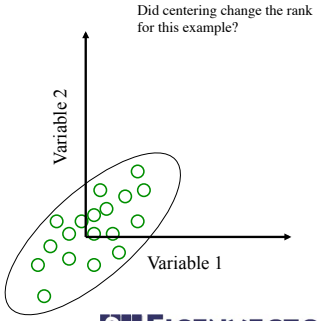
EIGENVECTOR RESEARCH INCORPORATED

Centering is an Axis Translation

- Geometry for 2 variables



Did centering change the rank for this example?



EIGENVECTOR RESEARCH INCORPORATED

Centering (general)

Centering as a projection

$$\mathbf{X}_{M\times N} \quad \text{and} \quad \bar{\mathbf{x}}^{\mathrm{T}} = \tfrac{1}{M}\mathbf{1}^{\mathrm{T}}\mathbf{X}$$

$$\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^{\mathrm{T}} = \mathbf{X} - \tfrac{1}{M}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{X} = \left(\mathbf{I} - \tfrac{1}{M}\mathbf{1}\mathbf{1}^{\mathrm{T}}\right)\mathbf{X}$$

Can center the data to something other than the mean

$\left(\mathbf{I} - \frac{1}{M}\mathbf{1}\mathbf{1}^{\mathrm{T}}\right)\mathbf{X}$	mean-centering
$\left(\mathbf{I} - \frac{1}{\mathbf{1}^{\mathrm{T}}\mathbf{w}}\mathbf{1}\mathbf{w}^{\mathrm{T}}\right)\mathbf{X},\ \mathbf{1}\mathbf{w}^{\mathrm{T}} \neq 0$	weighted mean-centering
$\mathbf{X} - \mathbf{1}\mathbf{w}^{\mathrm{T}}$	general center
$\mathbf{X}_{medcn} = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^{\mathrm{T}}_{median}$	median center

EIGENVECTOR RESEARCH INCORPORATED

Projection vs Subtraction

- example shows projection and subtraction give identical results

```
>> x = randn(10,4);
>> pm = eye(10)-ones(10,10)/10; % I - 11'/M
>> x1 = x-ones(10,1)*mean(x)     % x - 1x'   >> x2 = pm*x
x1 =                                        x2 =
  -0.4338  -0.4177   0.2691  -0.7587          -0.4338  -0.4177   0.2691  -0.7587
  -1.6669   0.4948  -1.3615   0.3312          -1.6669   0.4948  -1.3615   0.3312
   0.1241  -0.8193   0.6890   0.4568           0.1241  -0.8193   0.6890   0.4568
   0.2864   1.9522   1.5983   0.3531           0.2864   1.9522   1.5983   0.3531
  -1.1477  -0.3673  -0.7171   0.9314          -1.1477  -0.3673  -0.7171   0.9314
   1.1896  -0.1170   0.8327   0.3098           1.1896  -0.1170   0.8327   0.3098
   1.1879   0.8358   1.2287   0.8320           1.1879   0.8358   1.2287   0.8320
  -0.0389  -0.1717  -1.6190  -1.5613          -0.0389  -0.1717  -1.6190  -1.5613
   0.3260  -0.3266  -1.4662  -0.3786           0.3260  -0.3266  -1.4662  -0.3786
   0.1734  -1.0633   0.5459  -0.5156           0.1734  -1.0633   0.5459  -0.5156
```


EIGENVECTOR RESEARCH INCORPORATED

Centering Summary

- no offsets: $\mathbf{X} = \mathbf{T}\mathbf{P}^T$
- mean-centering: $\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{1}\bar{\mathbf{x}}^T$
- offset across mode 1: $\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{1}\boldsymbol{\mu}^T$
- offset across mode 2: $\mathbf{X} = \mathbf{T}\mathbf{P}^T + \boldsymbol{\mu}\mathbf{1}^T$
- offset across both modes: $\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mu\mathbf{1}\mathbf{1}^T$

The mean isn't the only, nor necessarily the best, factor to center to. The best depends on the objective and how the data manifest.

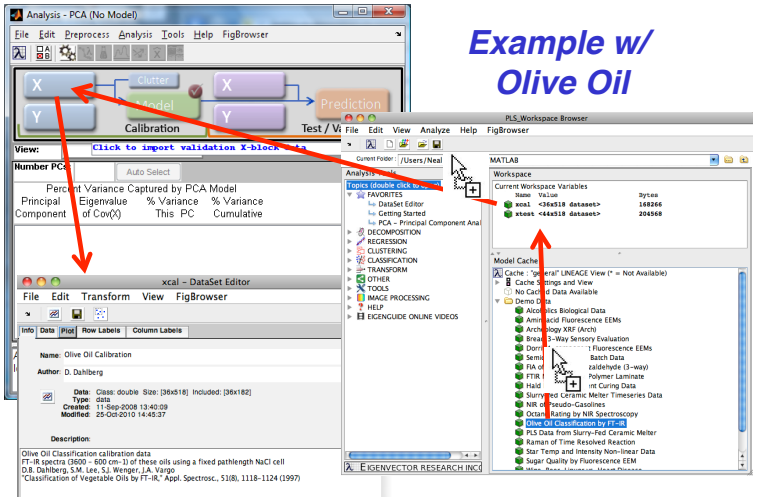
Centering increases fit to data, but probably not as much as an entirely unconstrained additional component.

$$\left\|\mathbf{X} - \mathbf{t}_1\mathbf{p}_1^T\right\| \geq \left\|\mathbf{X} - \left(\mathbf{t}_1\mathbf{p}_1^T + \mathbf{1}\bar{\mathbf{x}}^T\right)\right\| \geq \left\|\mathbf{X} - \left(\mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T\right)\right\|$$

10

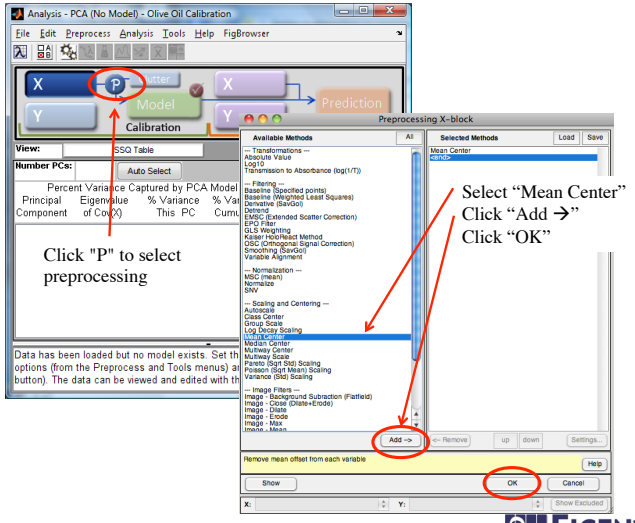
EIGENVECTOR RESEARCH INCORPORATED

Example w/ Olive Oil



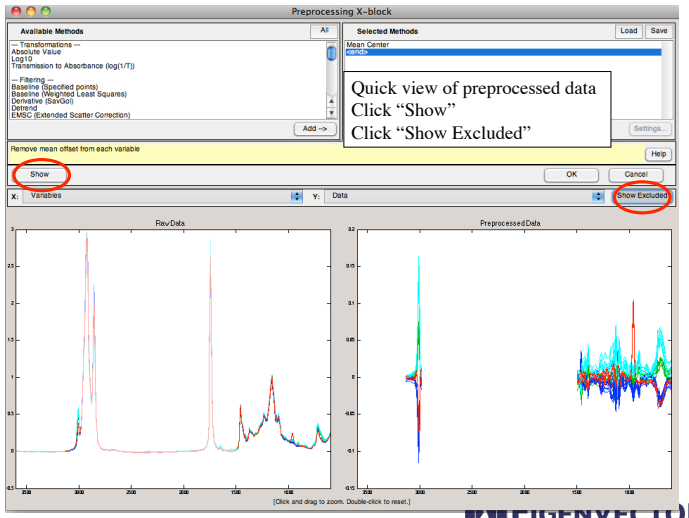
11

EIGENVECTOR RESEARCH INCORPORATED



13

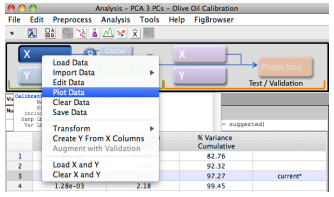
EIGENVECTOR RESEARCH INCORPORATED

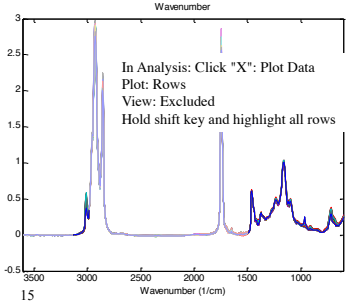


14

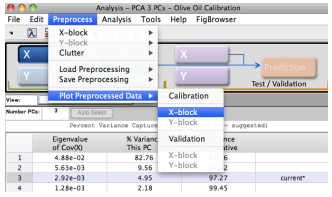
EIGENVECTOR RESEARCH INCORPORATED

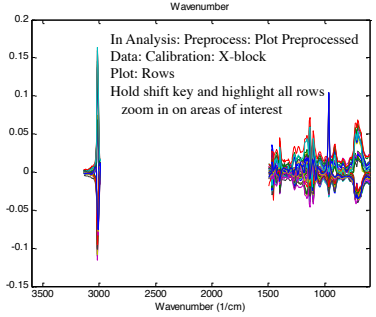
Plot the data





Plot the preprocessed data





Missing Data

- Many methods to replace missing data
  - the method used should not increase rank!
  - interpolation and other ad hoc procedures often work ~ok, but can increase rank
- Data can be replaced with values consistent with the overall data structure
  - often PCA is used (replaced entries have zero residual)
  - problem is knowing the number of factors
  - works if missing data are random (not systematic) and not too much missing (~<10%)

EIGENVECTOR RESEARCH INCORPORATED

Centering and Missing

Centering and missing
centering first and then fill in missing ~works if only a few missing, otherwise it adds rank

X is rank 1
even after centering, but not if data have missing values

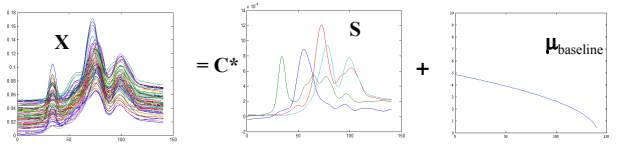
Alternative, fit the model
$||\mathbf{X}-\mathbf{TP}^T-\mathbf{1}\boldsymbol{\mu}^T||$ directly and replace with values consistent with the model mdcheck.m

	No missing	Missing
X	$\begin{bmatrix}1&2\\3&6\\2&4\\5&10\\9&18\end{bmatrix}$	$\begin{bmatrix}1&2\\3&6\\2&4\\?&10\\?&18\end{bmatrix}$
X_c	$\begin{bmatrix}-3&-6\\-1&-2\\-2&-4\\1&2\\5&10\end{bmatrix}$	$\begin{bmatrix}-1&-6\\1&-2\\0&-4\\?&2\\?&10\end{bmatrix}$

EIGENVECTOR RESEARCH INCORPORATED

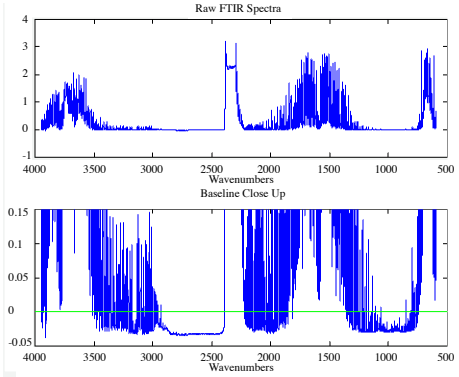
CLS Offsets (centering)

- Offsets for Classical Least Squares models
  - if offset is the same "spectrum" at different magnitudes it can be fit explicitly (and estimated using MCR)
    fit the model $||\mathbf{X}-\mathbf{CS}^T-\mathbf{1}\boldsymbol{\mu}^T||$ directly
  - if offset is different for each spectrum it must be removed using sample-specific pre-processing
    - baselining, other `baseline.m, baselinew.m, wlsbaseline.m`



EIGENVECTOR RESEARCH INCORPORATED

Sample-to-Sample Baseline



Baselines / backgrounds can manifest as offsets, sloping background, polynomial, or more complicated functions.

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

This type of **clutter** can inhibit predictive capability and make extraction of chemical information (e.g., via multivariate curve resolution) difficult.

EIGENVECTOR RESEARCH INCORPORATED

Background Subtraction

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

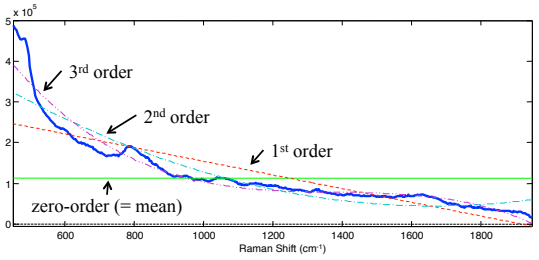
- **Detrend:** fit polynomial to *entire* spectrum
- **Selected-Points baselining:** fit polynomial to selected points in spectrum
- **Weighted Least-squares (a.k.a. asymmetric) baselining:** fit to *automatically* selected points on the bottom of the spectrum
- **Windowed:** Rolling Ball, Median, Minimum,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

Additional methods do baseline removal "within" the model (later…)

EIGENVECTOR RESEARCH INCORPORATED

Detrend

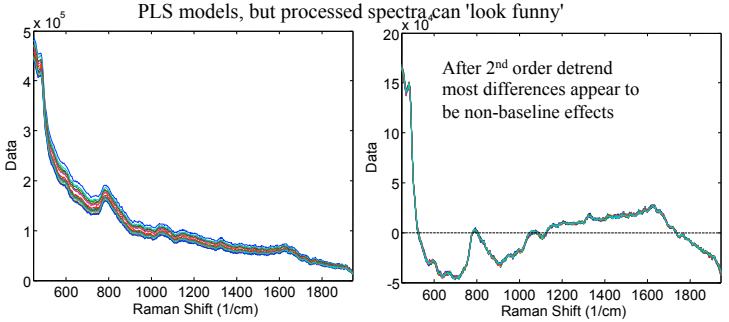
- Fit polynomial to entire spectrum
  - easy, but highly-influenced by non-baseline features



EIGENVECTOR RESEARCH INCORPORATED

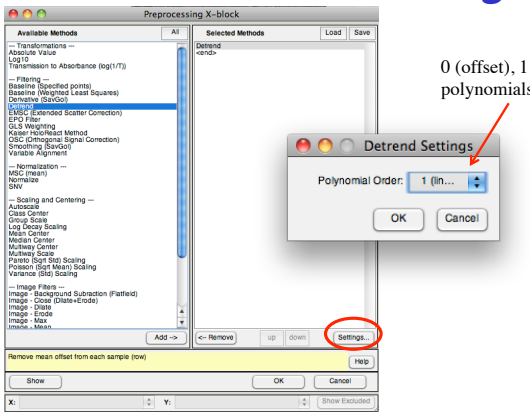
Detrend Example

- Raman Spectra of scattering sample
  - detrending good for exploratory analysis and can help with PLS models, but processed spectra can 'look funny'



EIGENVECTOR RESEARCH INCORPORATED

Detrend Settings

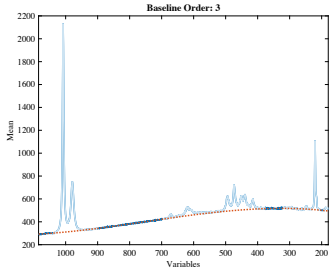


0 (offset), 1, 2, … higher order polynomials

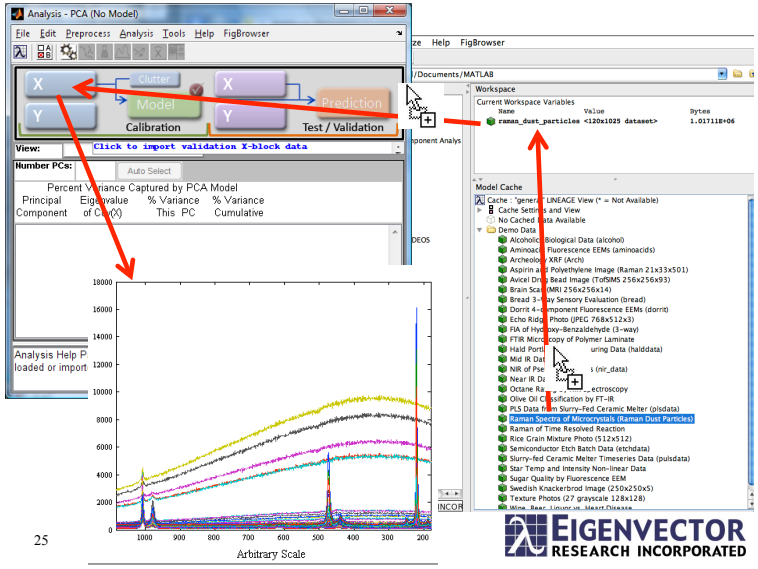
EIGENVECTOR RESEARCH INCORPORATED

Selected-Points Baseline

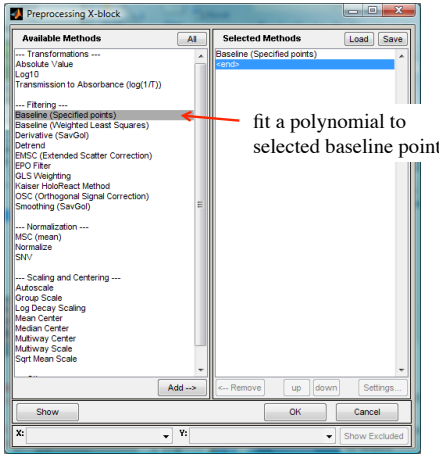
- Detrend based on points in spectrum known to be only baseline. Subtract the result from all channels.
  - good when zero points are known a priori



EIGENVECTOR RESEARCH INCORPORATED

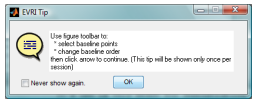


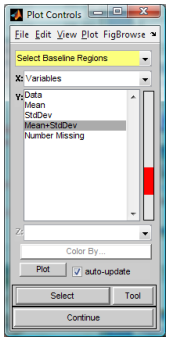
EIGENVECTOR RESEARCH INCORPORATED



fit a polynomial to selected baseline points

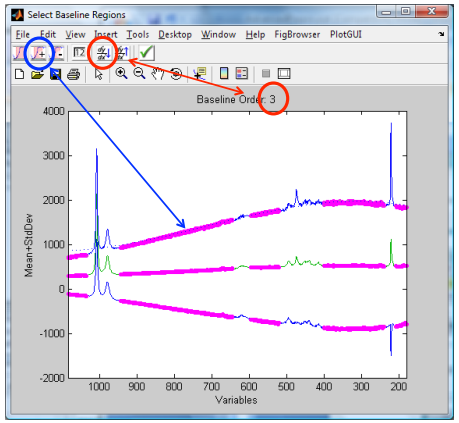
EIGENVECTOR RESEARCH INCORPORATED



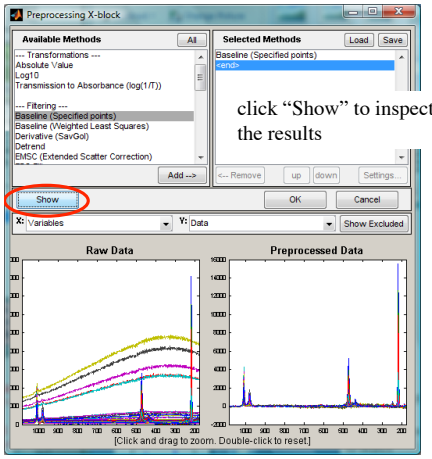


1. Click to select regions.
"+" is like holding shift key while selecting

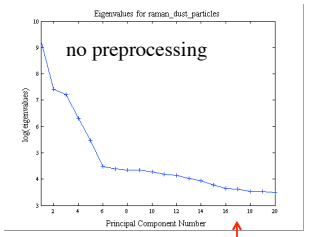
2. Increase polynomial order to fit to baseline points



EIGENVECTOR RESEARCH INCORPORATED

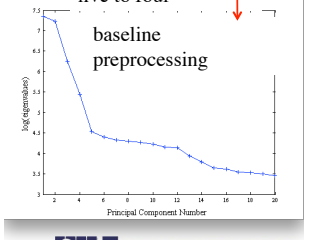


click "Show" to inspect the results



no preprocessing

rank dropped from five to four

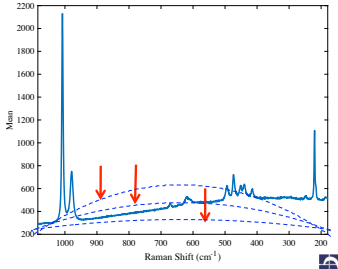


baseline preprocessing

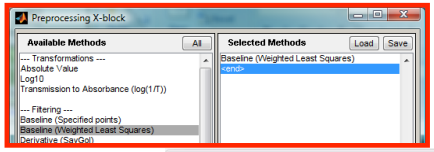
EIGENVECTOR RESEARCH INCORPORATED

Weighted Least-Squares Baselining

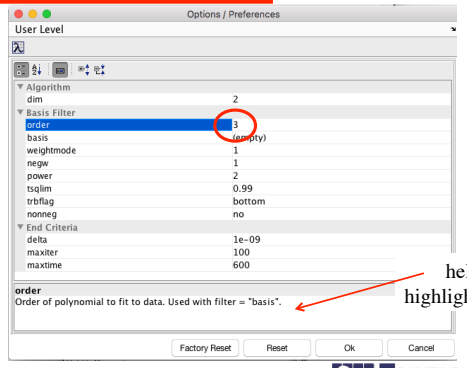
- Automatic selection of baseline points by fitting polynomial to the "bottom" (or "top") of the spectrum → asymmetric fit.
  - Starts with a fit to all points then de-weights points above the baseline (those with large positive residuals).
  - Iterates until only points w/in a defined tolerance on the 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



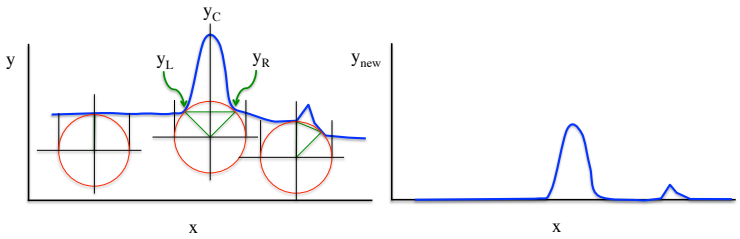
results similar to previous example



help for highlighted option

EIGENVECTOR RESEARCH INCORPORATED

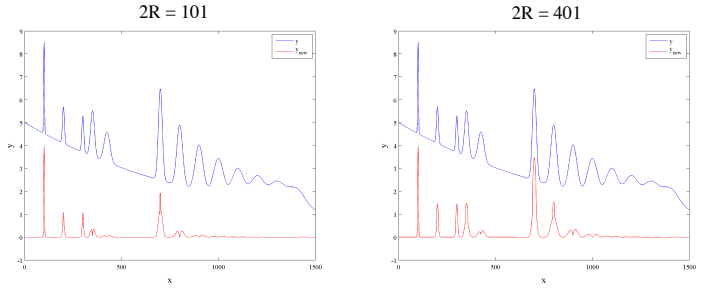
Rolling Ball Background Subtraction



$$y_{C,new} = y_C / ((y_L + y_R)/2)$$

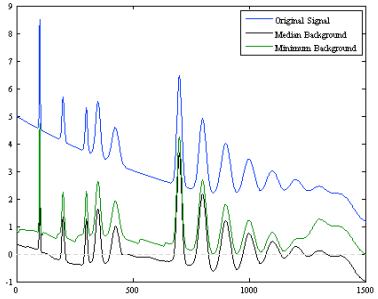
EIGENVECTOR RESEARCH INCORPORATED

Rolling Ball Example



EIGENVECTOR RESEARCH INCORPORATED

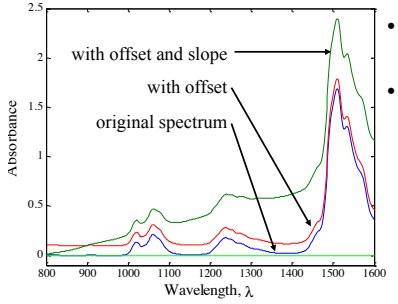
Other Background Subtraction Strategies



These appear to have fewer "peak artifacts" but there are still artifacts in the slope. Did it result in a drop in rank?

EIGENVECTOR RESEARCH INCORPORATED

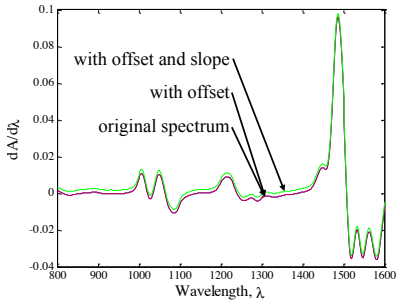
Savitzky-Golay and Filtering



- 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

```
load nir_data
x  = spec1.data(1,:);
x1 = spec1.data(1,:)+0.1;
x2 = spec1.data(1,:)+0.4+0.001*mncn(spec1.axisscale{2}')';
```

Savitzky-Golay and Filtering



multicomponent Beer's Law

$$\mathbf{x} = \mathbf{cS}^T$$

first derivative removes the offset

$$\mathbf{x} = \mathbf{cS}^T + \alpha\mathbf{1}^T$$

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

```
dx  = savgol(x, 15,2,1);
dx1 = savgol(x1,15,2,1);
dx2 = savgol(x2,15,2,1);
```

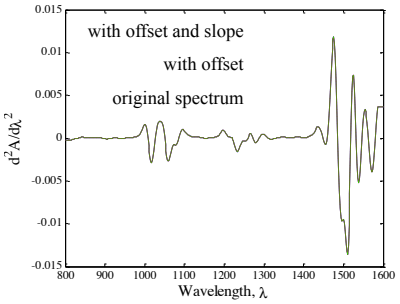
filter window width

polynomial order

derivative

EIGENVECTOR RESEARCH INCORPORATED

Savitzky-Golay and Filtering



multicomponent Beer's Law

$$\mathbf{x} = \mathbf{cS}^T$$

second derivative remove the offset and slope

$$\mathbf{x} = \mathbf{cS}^T + \alpha\mathbf{1}^T + \beta\boldsymbol{\lambda}$$

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

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

```
dx  = savgol(x, 15,2,2);
dx1 = savgol(x1,15,2,2);
dx2 = savgol(x2,15,2,2);
```

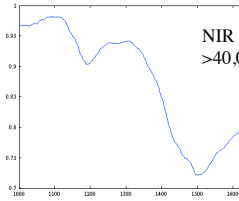
filter window width

polynomial order

derivative

EIGENVECTOR RESEARCH INCORPORATED

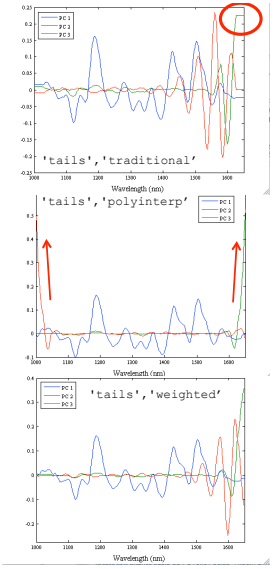
End-Effects in SavGol

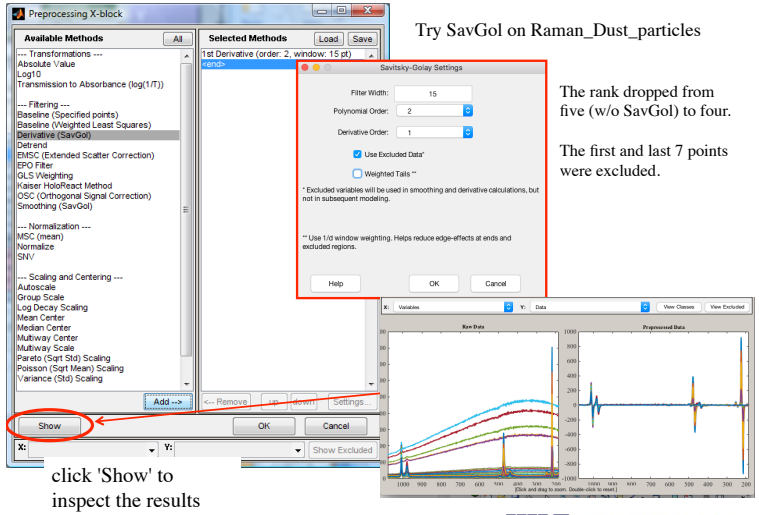


NIR of wheat gluten
>40,000 spectra

```
Principal    Eigenvalue     % Variance     % Variance
Component       of          Captured       Captured
 Number       Cov(X)         This  PC        Total
---------    ----------     ----------     ----------
savgol(x,25,2,2,struct('tails','traditional'))
      1        2.74e+02        84.58          84.58
      2        5.01e+00         1.55          86.12
      3        2.76e+00         0.85          86.98
savgol(x,25,2,2,struct('tails','polyinterp'))
      1        2.73e+02        78.76          78.76
      2        1.67e+01         4.83          83.59
      3        4.83e+00         1.40          84.98
savgol(x.data(i0,:),25,2,2,struct('tails','weighted'))
      1        2.73e+02        82.89          82.89
      2        7.50e+00         2.28          85.17
      3        2.97e+00         0.90          86.07
```


Data courtesy of Opotek, Inc., Carlsbad, CA, www.opotek.com





Try SavGol on Raman_Dust_particles

The rank dropped from five (w/o SavGol) to four.

The first and last 7 points were excluded.

click 'Show' to inspect the results

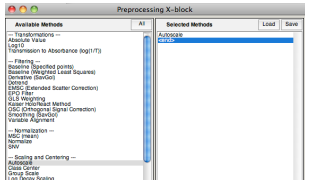
EIGENVECTOR RESEARCH INCORPORATED

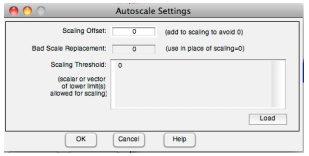
Least Squares Analysis

- Variables far from the origin have more influence on the model
  - numerically larger variables appear more important
  - *mean centering* can be used to reduce this effect but moves the model origin
  - if a force fit through zero is desired (e.g., often used with CLS-like models) it is important to remove offsets that add rank
    - » offset might not correspond to the mean of the data
- Variables with large deviations have more influence on the model
  - variance (squared deviations) is associated with "importance"
  - *autoscaling* - divide each (mean centered) variable by its standard deviation, result is variables with unit variance
  - other weighting and centering strategies might be more useful
    - » e.g., can use *a priori* information, such as noise level

Autoscaling

- **Autoscaling:** subtracts the column mean from each column and scales to unit variance
  - each variable converted to a T-statistic
  - result is that each variable has the same variance
  - not typically used with spectra
  - often used with engineering variables (different units)
- offset used to avoid "divide by zero"





Centering and Autoscaling Math

- Mean-centering `mncn(data)` $\mathbf{X}_{mncn} = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T$
- Autoscaling `auto(data)` $\mathbf{X}_{auto} = \left(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T\right)./\mathbf{1}\boldsymbol{\sigma}_x^T$
  - with an offset $\mathbf{X}_{auto} = \left(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T\right)./\mathbf{1}\left(\boldsymbol{\sigma}_x^T + \alpha\right)$

$\mathbf{X}_{auto} = \left(\mathbf{I} - \frac{1}{M}\mathbf{1}\mathbf{1}^T\right)\mathbf{X}\mathbf{W}^{-½}$

where $\mathbf{W}$ is a diagonal matrix with variance as the entries

$\mathbf{X}_{auto} = \left(\mathbf{I} - \frac{1}{M}\mathbf{1}\mathbf{1}^T\right)\mathbf{X}\left(\mathbf{W} + \alpha^2\mathbf{I}\right)^{-½}$

the offset α can be viewed as a ridging or regularization of $\mathbf{W}$

Weighting

- Auto-scaling can be viewed as a weighted approach where the weighting is the standard deviation of the variables
- Other weighting can be used
  - weight by noise in the variables with $\mathbf{W}$ diagonal
    - each variable has same S/N
  - $\mathbf{W}$ doesn't have to be diagonal
    - this will lead to generalized least squares approaches

Poisson Scaling

- For count data, the variance is expected to follow a Poisson distribution such that the variance is equal to the mean of the data.
  - often used in mass spectroscopy
  - in this case **W** is diagonal and equal to sqrt(mean(data))
  - M.R. Keenan, "Multivariate Analysis of Spectral Images Composed of Count Data," in *Techniques and Applications of Hyperspectral Image Analysis*, H. F. Grahn and P. Geladi, eds. (John Wiley & Sons, West Sussex, England), 89-126, 2007.
  - poissonscale.m

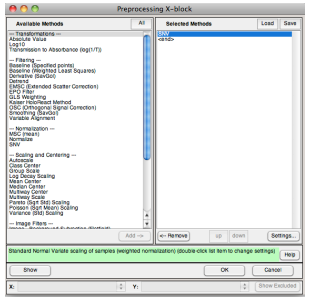
EIGENVECTOR RESEARCH INCORPORATED

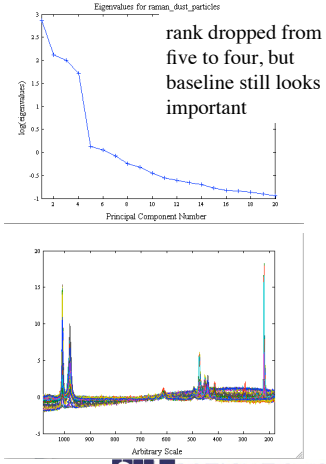
Removing the Row Magnitude

- Previous examples removed an offset. How is variance due to changing magnitude removed?
  - variable source or lighting magnitude
  - scattering effects
- **Standard Normal Variate (SNV)**: subtracts the row mean from each row and scales to unit variance
  - Autoscaling of the rows
- **Row Normalization**: removes magnitude
- Be aware that this can "blow up" noisy samples to have more variance

EIGENVECTOR RESEARCH INCORPORATED

SNV

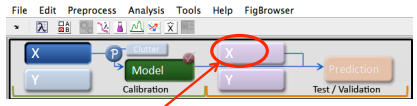




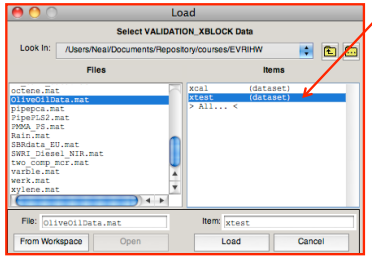
rank dropped from five to four, but baseline still looks important

see snv.m

EIGENVECTOR RESEARCH INCORPORATED



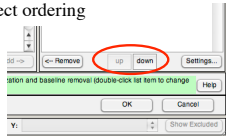
Click "X" to load **test data**



Example w/ Olive Oil

…\EVRIHW\OliveOilData.mat
xtest

1. Try mean-centering only and plot PC 2 vs PC 1.
2. Try SNV, then mean-centering. use "up" and/or "down" in the "preprocess" interface to ensure the correct ordering



EIGENVECTOR RESEARCH INCORPORATED

Normalization

- ***Row*** normalization is used to remove magnitude information from each sample
  - ***column*** normalization was used to put variables on similar scales (autoscaling, other scaling)
  - there are many ways to normalize
  - often used with spectra to remove scattering/pathlength differences observed in the measurements
- Standard normal variate (SNV)
  - autoscales rows (centering and scaling ~similar to centering and 2-norm)

see `normaliz.m`

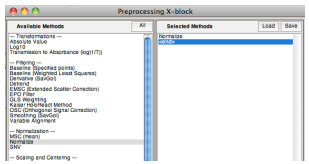
EIGENVECTOR RESEARCH INCORPORATED

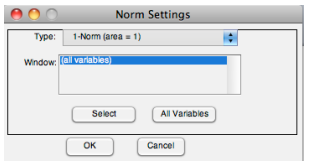
Normalization

- 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

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

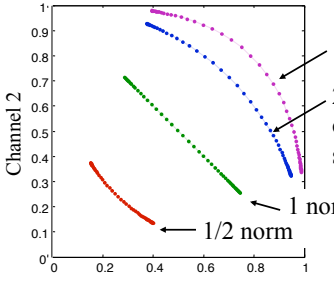




EIGENVECTOR RESEARCH INCORPORATED

Normalization

- Normalize each row / spectrum (which p?)



p norm

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

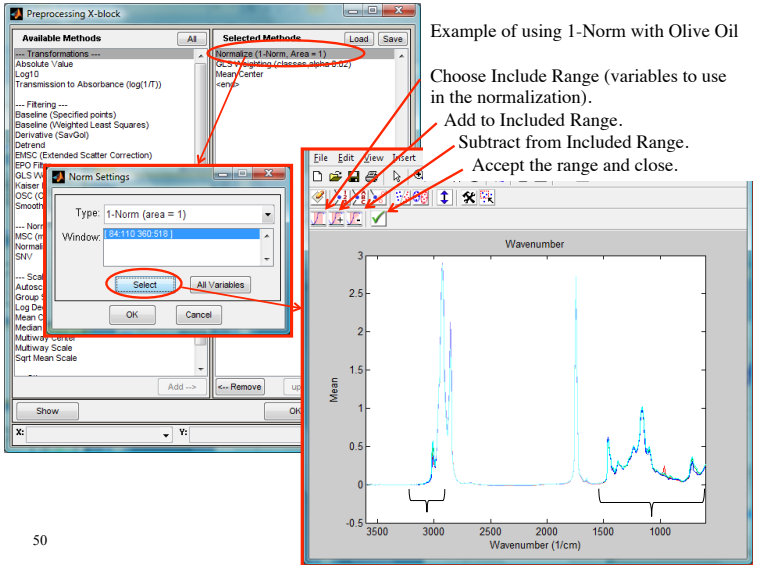
3 norm

2 norm: constrains rows to a spherical surface

1 norm: constrains rows to a plane

Gallagher, NB, Shaver, JM, Martin, EB, Morris, J, Wise, BM, Windig, W, "Curve resolution for images with applications to TOF-SIMS and Raman", *Chemometr. Intell. Lab.*, **73**(1), 105–117 (2003). See **Section 3.3**.

EIGENVECTOR RESEARCH INCORPORATED



Example of using 1-Norm with Olive Oil

Choose Include Range (variables to use in the normalization).

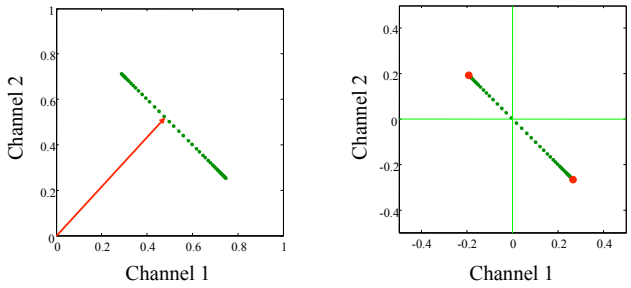
Add to Included Range.

Subtract from Included Range.

Accept the range and close.

1 Norm

- Mean centering the 1 norm spectrum drops the rank
  - samples with small norm not used



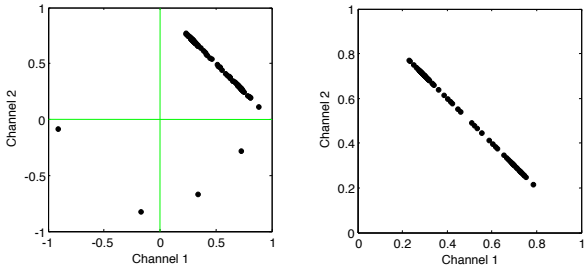
51

EIGENVECTOR RESEARCH INCORPORATED

What Happens with Noise?

- Remove samples with norm <0.5
  - or add a small offset

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



52

EIGENVECTOR RESEARCH INCORPORATED

1 Norm + Mean-Center → Endmember Extraction

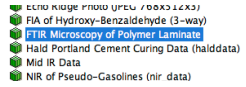
- Points at the vertices of the polygon correspond to the most "pure" samples (or variables)
  - if selectivity is high enough the vertices correspond to pure analyte contributions (or spectra)
  - this is a good visualization tool and initial guess for self-modeling mixture analysis (a.k.a. multivariate curve resolution or endmember extraction)

53

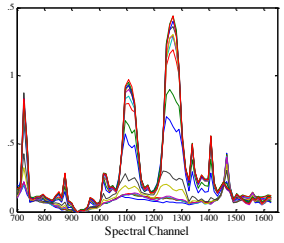
EIGENVECTOR RESEARCH INCORPORATED

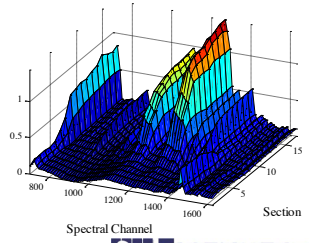
E.g., FTIR microscopy of three-layer polymer laminate

- Polyetheylene, isophtalic polyester (presence originally unknown) and polyethylene terephthalate.
- Laminate is 240 μm thick with inner IPE layer 2-3 μm < the 10 μm spatial resolution
  - middle layer has contributions from all three analytes
- Scanned at seventeen points across the layers



…/PLS_Toolbox/dems/FTIR_microscopy.mat



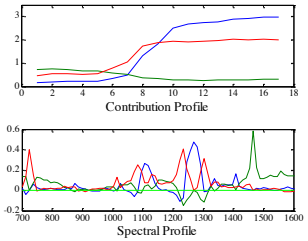


54

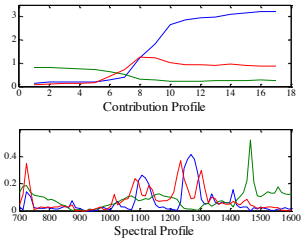
EIGENVECTOR RESEARCH INCORPORATED

FTIR_Microscopy Results

results of selecting vertices of the rows ("purest" spectra)



results after using non-negatively constrained alternating least squares



J. Guilment, S. Markel, W. Windig, Infrared chemical micro-imaging assisted by interactive self-modeling multivariate analysis, *Appl. Spectr.*, **48**, 1994, 320-326.
W. Windig, S. Markel, Simple-to-use interactive self-modeling mixture analysis of FTIR microscopy data, *J. Molecular Structure*, **292**, 1993, 161-170.

EIGENVECTOR RESEARCH INCORPORATED

Scatter / Signal Correction

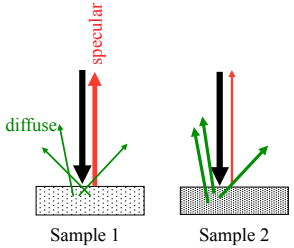
- Multiplicative Scatter Correction (MSC)
  - Attempts to remove offset *and* row magnitude variability
  - Result is less signal related to scattering artifacts and more signal related to analyte(s) of interest
  - based on classical least squares (CLS) model
  - Geladi P, MacDougall D, Martens H., *Appl. Spectrosc.*, **39**(3), 491-500 (1985)

EIGENVECTOR RESEARCH INCORPORATED

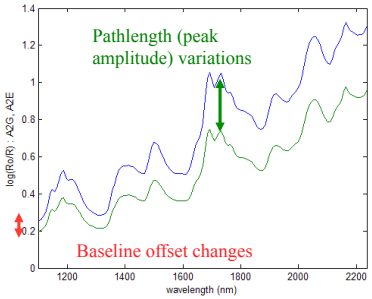
Scattering Effects in Reflectance

Scattering effects can be caused by variations in:

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

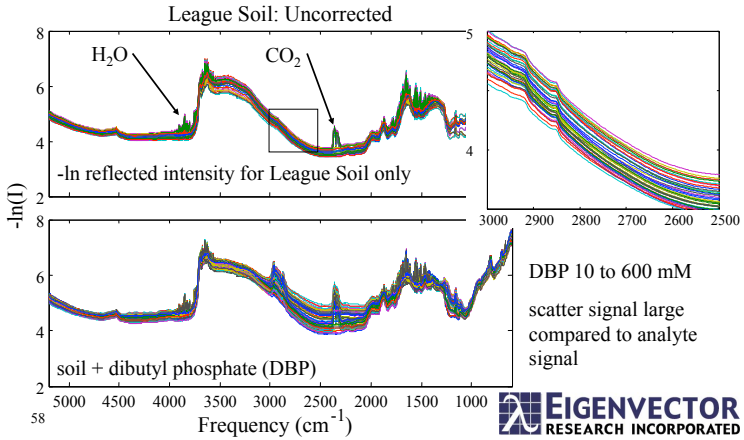


Scattering effects can manifest in the measurements as:



EIGENVECTOR RESEARCH INCORPORATED

Why is MSC Necessary?



EIGENVECTOR RESEARCH INCORPORATED

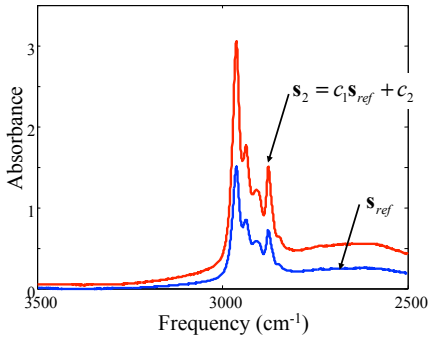
Multiplicative Scatter Correction (MSC)

MSC models scatter with an offset and slope correction

$$\mathbf{s}_2^T = \mathbf{c}\begin{bmatrix}\mathbf{s}_{ref} & \mathbf{1}\end{bmatrix}^T$$

where $\mathbf{c}_{1x2}$

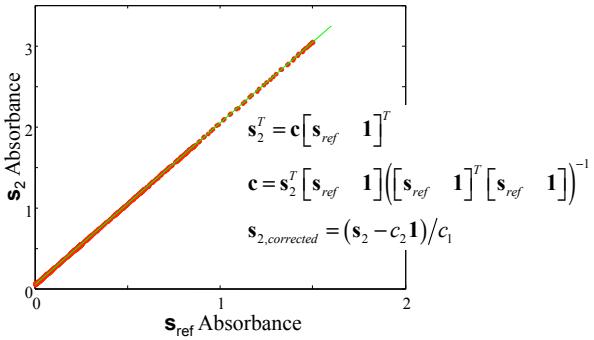
This is a classical least squares (CLS) model



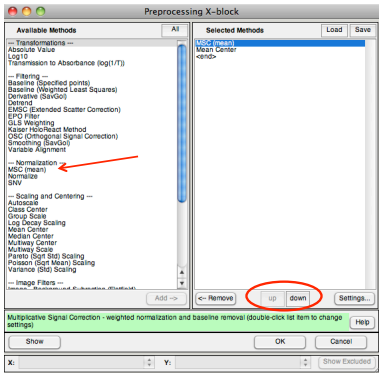
EIGENVECTOR RESEARCH INCORPORATED

MSC

Regress new measured spectrum $\mathbf{s}_{2,Nx1}$ onto reference spectrum $\mathbf{s}_{ref,Nx1}$ (often the mean spectrum of a data set).

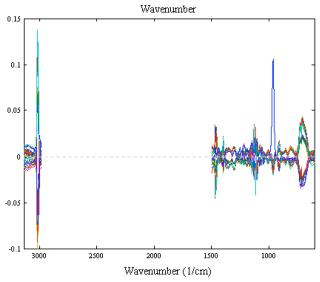


EIGENVECTOR RESEARCH INCORPORATED



try with the olive oil data

after adding MSC (mean) and Mean Center
use "up" and/or "down" to ensure the correct ordering



EIGENVECTOR RESEARCH INCORPORATED

Pre-process Summary to Now

- Centering
  - May be more complicated than simple mean-centering
  - Did it add or remove rank? Usually want to remove rank.
- Baselining
  - detrend, selected points, asymmetric least squares
  - SavGol
    - didn't remove relevant signal some preprocessing will (beware!)
- Column scaling (autoscaling)
- Row scaling (SNV, normalization)
  - Column and Row scaling introduces the concept of general centering and weighting …
- MSC attempts to remove offset and row magnitude variability

EIGENVECTOR RESEARCH INCORPORATED

Why Pre-process?

- Pre-processing depends on the analysis objective and data type.
  - The objective of data pre-processing is to remove artifacts, interferences and non-linearity in the data.
  - Sensor systems often include measurement artifacts that hinder analysis and negatively impact the quality of the results.
  - Math is cheaper than physics but it a'int magic.
  - Preprocessing isn't a silver bullet for bad data.
  - Knowledge of the opportunities and limitations of the math leads to better data through instrument design and data acquisition methodology …
  - and better results in the subsequent data analysis and fulfilling the analysis objective.

Data Analysis Objectives

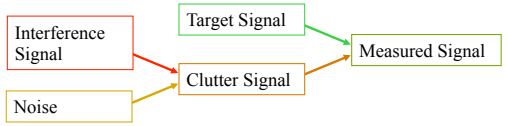
- Be clear on the data analysis objective
  - exploratory
    - how do measurements vary and why?
    - how data are preprocessed tells how results should be interpreted
  - process monitoring/control, quality control
    - multivariate statistical process control typically monitors how data vary wrt a mean (or moving mean)
  - detection
    - is new signal present? relative to 0 or clutter mean?
  - classification
    - some variables are better at discriminating some classes better than others (utilize different scaling for discriminating variables?)
  - quantification
    - mean-center allows an offset, otherwise it's a force fit through zero

Physics and Chemistry

- How and why preprocessing depends on the physics and chemistry of the system
- … and how the data manifest
  - is it noisy, non-linear, stationary, cluttered, multi-block, multi-way?
- Data analysis is not a limited access highway. It is more like a worn path in the dirt. What is learned at each step may cause us to return to the beginning to test our hypotheses.
  - Preprocess any way you want but you must interpret your results accordingly …
  - and ask if you achieved your overall objective

Measured Signal

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



  - Use physics to create a linear relationship
    - non-linearity w/in X-block adds factors
    - non-linearity between X- and Y-blocks adds error

Sources of Clutter

- Systematic background variability
  - 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.
  - in the system being sensed
    - e.g., T, P changes, variable sample matrix, "dark current"
    - use pre-processing or different sensing strategies
  - due to physics of instrument
    - e.g., drift, instrument changes, variable baseline or gain
    - try pre-processing but good instrument design and operational practices preferable
- Non-systematic random noise
  - homoscedastic, heteroscedastic

Matrix Rank

- Matrix rank is an important concept
  - pre-processing generally tries to reduce rank associated with clutter while retaining rank associated with signal
  - … quick review of rank via the PCA model

PCA Math Summary

- For a data matrix $\mathbf{X}$ with M samples and N variables (generally assumed to be mean centered and properly scaled), the PCA decomposition is

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

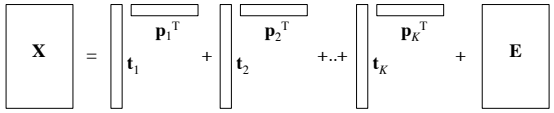
  Where $R \leq \min\{M,N\}$, and the $\boldsymbol{t}_k\boldsymbol{p}_k^T$ pairs are ordered by the amount of variance captured.
- Generally, the model is truncated to K PCs that capture the systematic variance in the data set, leaving some small amount of variance in a residual matrix $\mathbf{E}$:

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

- where $\mathbf{T}$ is MxK and $\mathbf{P}$ is NxK.

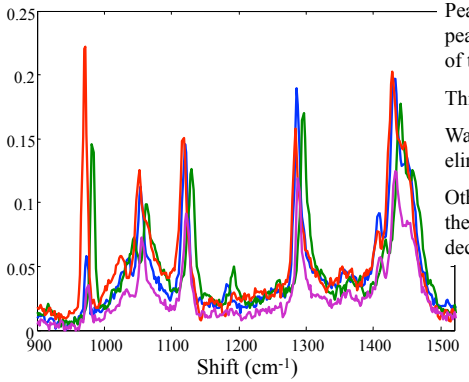
PCAENGINE, PCA

Matrix Rank and the Bilinear Model



- Sources of matrix rank
  - chemical signal (signal of interest; often called the pseudo-rank)
  - clutter (interference signal not of interest and noise)
    - interference that inhibits the ability to detect, classify, quantify
    - want to remove clutter-based rank and keep the chemical source of rank

Clutter Example (peak shift)



Peak instability (i.e., shift in peak location) increases the rank of the data.

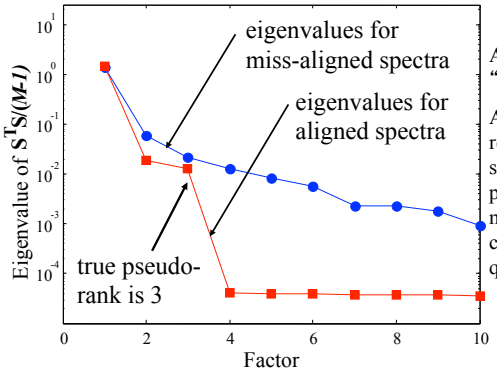
This additional rank is "clutter".

Wavelength calibration can eliminate this type of clutter.

Other methods attempt to model the shifts as a part of the decomposition (e.g., MCR).

EIGENVECTOR RESEARCH INCORPORATED

Peak Shift Example



Aligned spectra have less "clutter" and lower rank.

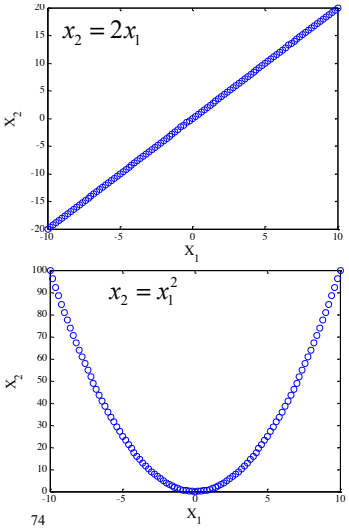
Additional "clutter" rank often results in a loss of net analyte signal and can degrade the performance of regression models ➔ inhibits detection, classification, and quantification.

EIGENVECTOR RESEARCH INCORPORATED

Data Transformation

- Within an X-block
  - PCA assumes relationships between variables are linear
- Between and X- and Y-block
  - PCR, PLS assume relationship is linear
- If possible, non-linear data should be converted to a linear form (e.g., use known physics of the system)
- Examples:
  - I/I_0, transform with log
  - reaction rates a $e^{-1/T}$, transform with log
  - pipe flow a $\Delta P^{4/7}$ (turbulent flow)

EIGENVECTOR RESEARCH INCORPORATED



PCA Linear vs Non-Linear

% Variance Captured by PCA Model

PC Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
Linear system			
1	1.72e+002	100.00	100.00
Non-linear system			
1	2.10e+003	98.39	98.39
2	3.43e+001	1.61	100.00

EIGENVECTOR RESEARCH INCORPORATED

Linear vs Non-Linear

- Non-linearity between the variables adds rank
- For signal-to-noise > 1 (signal-to-clutter > 1)
  - The signal of interest is primarily in the big Principal Components PCs (1,2,…,*K*)
  - The clutter is primarily in the smaller Principal Components (*R*-*K*+1, *R*-*K*+2…, *R*)
    - More on PCA shortly
  - Linearization has the potential of bringing redundant signal "closer to the top" (into the big PCs)
    - better signal averaging

CLS vs. PCA (Bilinear Models)

- Classical Least Squares
  - $\mathbf{X} = \mathbf{C}\mathbf{S}^T + \mathbf{E}$
  - $\mathbf{X}_{MxN}$
  - $\mathbf{C}_{MxK}$; oblique, chemically meaningful
  - $\mathbf{S}_{MxK}$; oblique, chemically meaningful
  - $\mathbf{E}_{MxN}$
  - multi-component Beer's Law model , **E** ~not minimized (e.g., due to constraints)
- PCA
  - $\mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$
  - $\mathbf{X}_{MxN}$
  - $\mathbf{T}_{MxK}$; orthogonal, ~not chemically meaningful
  - $\mathbf{P}_{MxK}$; orthogonal, ~not chemically meaningful
  - $\mathbf{E}_{MxN}$
  - captures maximum variance, **E** minimized

Why is Clutter Bad?

- Show with CLS: multi-component Beer's law

$$\mathbf{X} = \mathbf{c}_1\mathbf{s}_1^T + \mathbf{c}_2\mathbf{s}_2^T + \ldots \mathbf{c}_K\mathbf{s}_K^T = \mathbf{C}\mathbf{S}^T$$

$\mathbf{X}$ M by N data matrix (noise-free)

$\mathbf{c}_i$ M by 1 vector (concentration)

$\mathbf{s}_i$ N by 1 vector (spectrum)

K number of chemical analytes present $i = 1,\ldots,K$

Clutter Reduces Net Analyte Signal

If the measured signal is $\mathbf{x} = \mathbf{c}^T\mathbf{S}^T + \mathbf{e}^T$ then define the Net Analyte Signal ***Vector***, NAS at unit concentration as

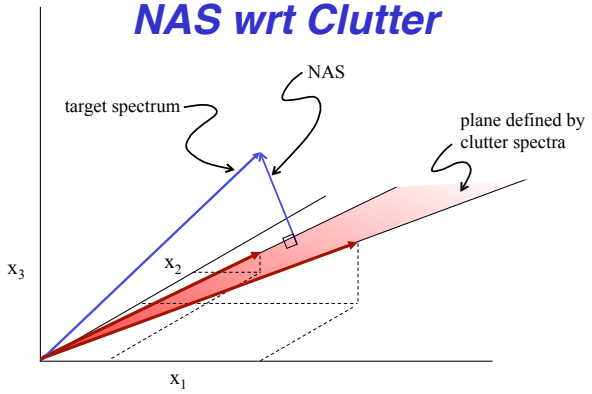
$$\text{NAS} = [\mathbf{I} - \mathbf{S}_{-i}(\mathbf{S}_{-i}^T\mathbf{S}_{-i})\,\mathbf{S}_{-i}^T]\mathbf{s}_i$$

where $\mathbf{S}_{-i}$ is S with the i^{th} column removed.

NAS is the portion of spectrum $\mathbf{s}_i$ unique to analyte i and orthogonal to all other factors in $\mathbf{S}_{-i}$, and S/N ~ |NAS|

Adding clutter tends to add something in $\mathbf{S}_{-i}$ that is parallel to $\mathbf{s}_i$ thus lowering NAS (and increasing the estimation error).

NAS wrt Clutter



Every time an interference is added (new clutter) there is a risk of reducing NAS.
Estimation error ~1/|NAS|

EIGENVECTOR RESEARCH INCORPORATED

Example of NAS Decreasing

- Imagine there are five potential analytes and the first is the target of interest. NAS is a vector.

$$NAS = \left(\mathbf{I} - \mathbf{S}_{-i}\mathbf{S}_{-i}^{T}\mathbf{S}_{-i}\mathbf{S}_{-i}^{T}\right)\mathbf{s}_{i}$$

$\mathbf{s}_i$ is the i^{th} column of $\mathbf{S}$

$\mathbf{S}_{-i}$ is $\mathbf{S}$ with the i^{th} column removed

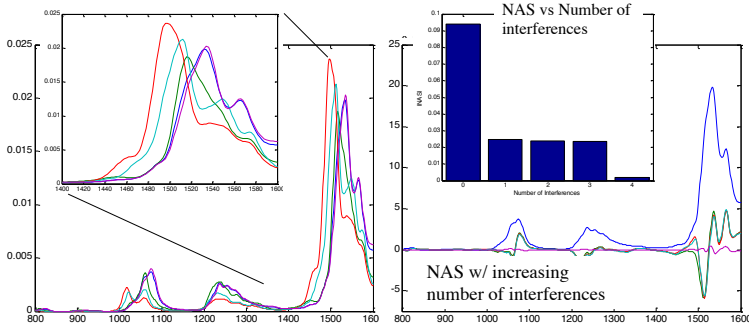
```
>> load nir_data
>> whos
  Name          Size              Bytes  Class

  conc         30x5                7408  dataset
  readme        7x67                938  char
  spec1        30x401            109008  dataset
  spec2        30x401            109008  dataset
>> pspec = conc.data\spec1.data;  % S
>> plot(spec1.axisscale{2},pspec)
```


EIGENVECTOR RESEARCH INCORPORATED

Interferences Decrease NAS

```
>> snas = pspec;
>> for i=2:5, snas(i,:) = pspec(1,:)-(pspec(1,:)/pspec(2:i,:))*pspec(2:i,:); end
```



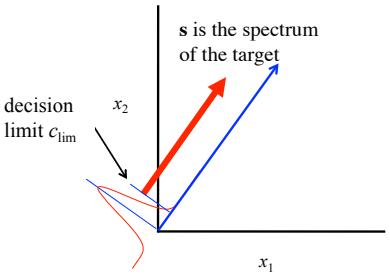
EIGENVECTOR RESEARCH INCORPORATED

How to Account for Clutter?

- Follow through with the CLS example and show how we might deal with clutter using the extended mixture model (ELS).
  - Target detection example
- Extended multiplicative scatter correction (EMSC)
  - Combines the extended mixture model and multiplicative scatter correction
  - Introduce, multiplicative scatter correction (MSC) first

EIGENVECTOR RESEARCH INCORPORATED

Target Detection for a Single Target



Measured signals > decision limit in the direction of the target is considered a detection event.

The measured signal is modeled as a linear multiple of the target signal.

$\mathbf{x} = c\mathbf{s} + \mathbf{e}$

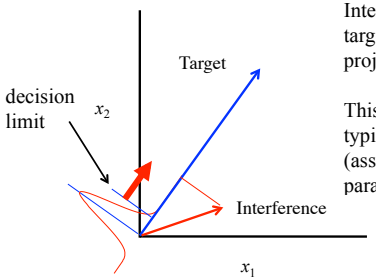
$c > c_{lim}$ is a detection

This is the CLS model.

- Detection on the target is often based on a non-negativity constraints.
- Present model assumes only a single target.
- What to do with signal from interferences and / or other targets?

EIGENVECTOR RESEARCH INCORPORATED

Target Detection with an Unknown Interference



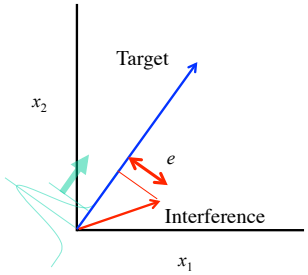
Interferences signal that overlaps the target spectrum results in a positive projection on the target.

This would be a false alarm, but typically more information is available. (assuming the interference isn't exactly parallel to the target)

- Presence of an unknown interference can result in false alarms.

EIGENVECTOR RESEARCH INCORPORATED

Target Detection with an Unknown Interference



- Unknown interference typically results in "no-calls".

The difference between the interference signal and the target spectrum provides information about the signal.

$\mathbf{e} = \mathbf{x} - c\mathbf{s}$

$Q = \mathbf{e}^T\mathbf{e}$

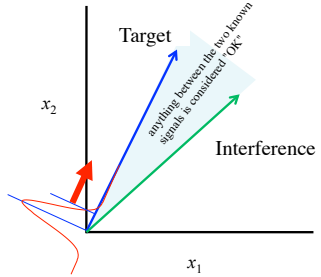
If the measured signal looks like target Q is small and there is good confidence in the detection.

If the measured signal does not look like target, Q is large and there is poor confidence in the detection i.e., this is a "no-call".

How can interferences be accounted for so that detection can be made with high confidence?

EIGENVECTOR RESEARCH INCORPORATED

Target Detection Accounting for an Interference with the Extended Mixture Model



Including an interference explicitly in the model allows the detector to account for interference signal.

$\mathbf{e} = \mathbf{x} - \mathbf{cS}$

$\mathbf{e} = \mathbf{x} - [c_1\ c_2][\mathbf{s}_1\ \mathbf{s}_2]$

where $\mathbf{s}_1$ is the target spectrum and $\mathbf{s}_2$ is the interference spectrum

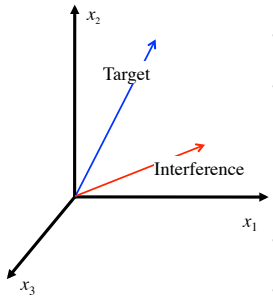
$\mathbf{c}_1 > c_{lim,1}$ is a detection

This is the extended mixture model (Extended Least Squares).

- Multiple targets and interferences can be accounted for (but there's a limit).
- Non-negativity forces the signal to lie on, or between, the target and interference.
- Detection statistics for targets are treated independently from interferences .

EIGENVECTOR RESEARCH INCORPORATED

Target Detection (summary)



- The model for target detection is based on a Linear Mixture model
- Non-negativity incorporates *a priori* knowledge of the physics.
- Projections onto the target vectors
  1. provides the contribution of each target to the measured signal where contributions determine the mix of targets (can be used for classification)
  2. the combined distance along the targets (T^2) provides a measure of distance from the null
- New measurements that lie within the plane of the targets are considered nominal signals.
- New measurements that lie off the plane of the targets are considered unusual signals. Unusual signals occur due to sensor problems or interferences – and correspond to no-calls. These signals are defined by limits on Q (sum-squared-residuals).

EIGENVECTOR RESEARCH INCORPORATED

Extended and Inverse MSC

- Extended multiplicative scatter correction (EMSC)
  - based on CLS and the extended mixture model
- Inverse SC and Extended Inverse SC
  - based on inverse least squares (ILS) model
  - Martens H, Stark E., *Journal of Pharmaceutical 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).

EIGENVECTOR RESEARCH INCORPORATED

Objective

- Remove
  - scattering artifacts
  - atmosphere analytes (H_2O and CO_2)
- Retain
  - reference soil signal
  - analyte signal [dibutyl phosphate (DBP) – organophosphorous]
- However, scattering and atms are more complicated than simple offset and gain
  - extended least squares model

EIGENVECTOR RESEARCH INCORPORATED

Windowed Scatter Correction

- Scatter correction is typically applied to entire frequency range, BUT….
- Windowing (piece-wise) correction is possible!
  - Requires more parameters, but is more flexible
    - If offset/multiplicative factors are NOT constant with wavelength
  - Isaksson T, Kowalski B., *Appl. Spectrosc.*, **47**(7), 702-709 (1993).
  - Blank TB, Sum ST, Brown SD, Monfre, SL., *Anal. Chem.*, **68**(17), 2987–2995 (1996).

EIGENVECTOR RESEARCH INCORPORATED

Extended MSC

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

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

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

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

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

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

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

EIGENVECTOR RESEARCH INCORPORATED

EMSC

can add spectra of known ***target analyte*** $\mathbf{S}_{A,NxJ}$

polynomial can be of order K-1

can add spectra of known ***interference*** $\mathbf{Q}_{NxL}$.

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

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

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

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

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

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

EIGENVECTOR RESEARCH INCORPORATED

"Reality Check": How to get Q?

- **Q** is a sub-space that spans scatter
  - **Q** spans the clutter
  - measure multiple reflectance spectra of soil samples that do not contain analyte → $\mathbf{X}_Q$
  - define reference spectra as mean of $\mathbf{X}_Q$ → $\mathbf{s}_{ref}$
  - center $\mathbf{X}_Q$ to $\mathbf{s}_{ref}$ → $\mathbf{X}_{Qm}$
  - perform PCA on centered data → $\mathbf{X}_{Qm}=\mathbf{T}\mathbf{Q}^T + \mathbf{E}$
    - use the big eigenvalues to get **Q**
  - use the loadings $\mathbf{Q}_{NxL}$ to characterize scatter

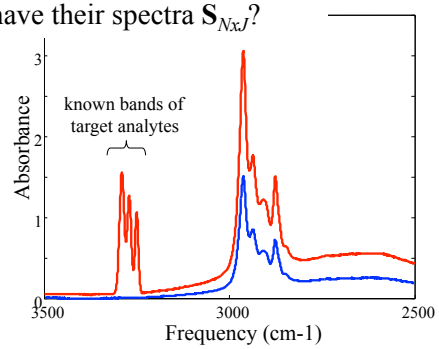
EIGENVECTOR RESEARCH INCORPORATED

Extended MSC Weighted

What if we know channels that should have target analytes but we don't have their spectra $\mathbf{S}_{NxJ}$?

If we do nothing the target spectra will bias the regression and provide poor correction.

Weighted least squares can be used to de-weight these channels in the MSC regression step.



EIGENVECTOR RESEARCH INCORPORATED

Weighted EMSC

- use a diagonal weighting matrix $\mathbf{W}_{NxN}$ to de-weight channels where analyte is present
  - weights are 0 where analyte is expected to be present

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

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

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

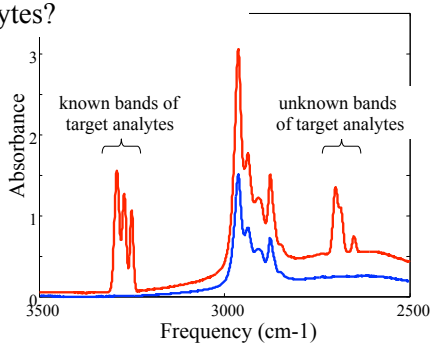
EIGENVECTOR RESEARCH INCORPORATED

Robust EMSC

What if we don't know all the channels that should have target analytes?

If we do nothing the target spectra will bias the regression and provide poor correction.

Robust least squares (treats these points like "outliers") can be used to de-weight these channels in the regression step.



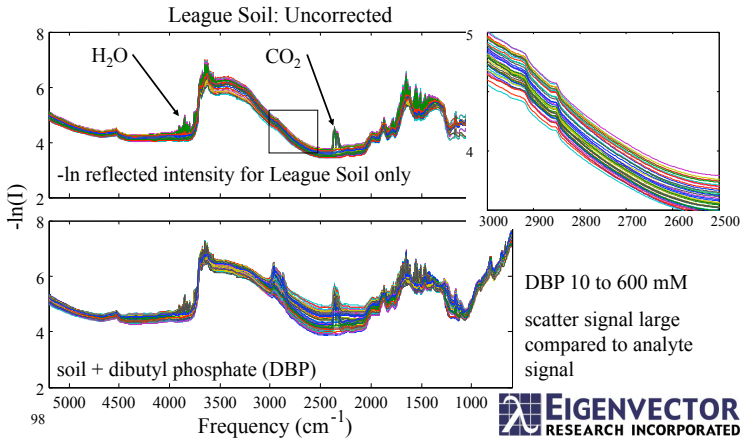
EIGENVECTOR RESEARCH INCORPORATED

Soil/Analyte Samples

- League Soil (44% clay, 42% silt, 14% sand)
- Quincy Soil (7% clay, 17% silt, 76% sand)
- Analyte: Dibutyl phosphate in 2-Methyl Butane
  - 0, 10-600 mM dripped onto soil sample
  - 2 MB highly volatile, evaporates quickly
  - measure spectra w/ and w/o dry-N_2 purge
  - sample (DBP concentration) randomized

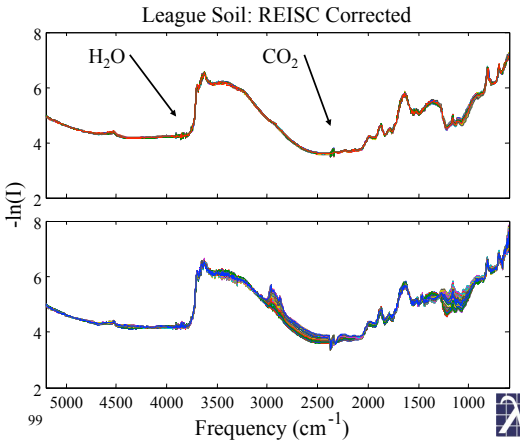
EIGENVECTOR RESEARCH INCORPORATED

Scatter on League Soil



EIGENVECTOR RESEARCH INCORPORATED

Robust EISC for League Soil

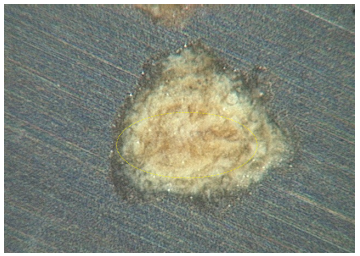


DBP 10 to 600 mM

scatter signal much smaller compared to analyte signal

EIGENVECTOR RESEARCH INCORPORATED

MCR w/ Clutter Example
Using PCs for the ELS in MCR



Where are the sugar and protein in a feed pellet?

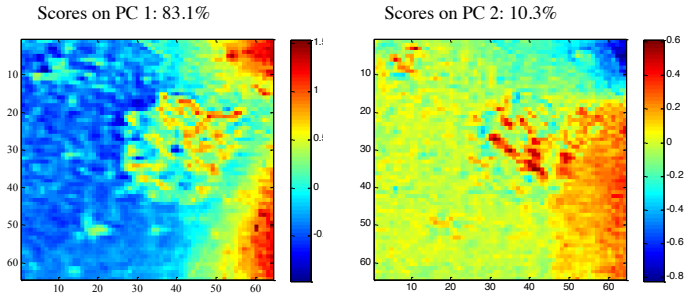
Embed a pellet in epoxy, section, and polish. Scratches are evident and are a source of significant clutter. Confounds the analysis so that chemical information is "smeared out" in the image.

Thanks to Sean Smith and Janiece Hope of Cargill, Inc., Global Food Research, Scientific Resources for the image data.

FTIR reflection image ~400 microns square

EIGENVECTOR RESEARCH INCORPORATED

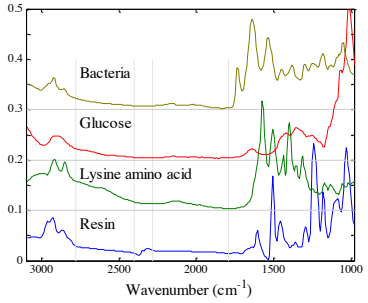
PCA



PCA Scores for mean-centered data. Not directly chemically meaningful, so let's try multivariate curve resolution.
How to deal with signal from scratches?

EIGENVECTOR RESEARCH INCORPORATED

Initial Estimates for Spectra



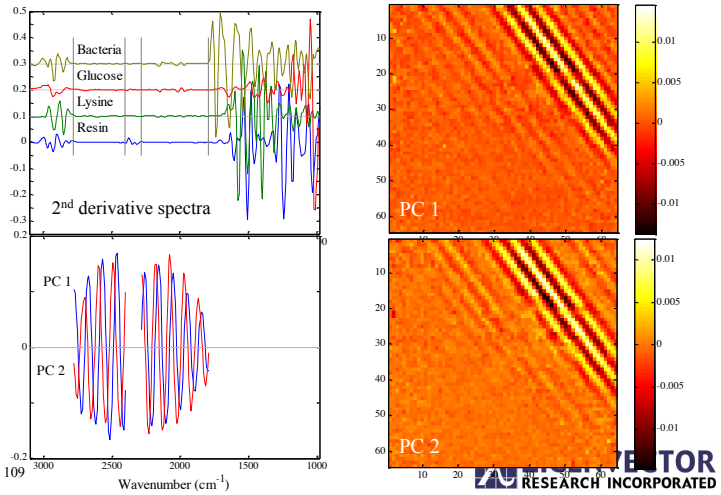
Step 1: isolate the clutter

regions used with 2nd derivative spectra to estimate spatial contributions of scratch features

can this information about the scratches be used to improve estimates of chemical contributions to the image in MCR?

EIGENVECTOR RESEARCH INCORPORATED

Scratch Features



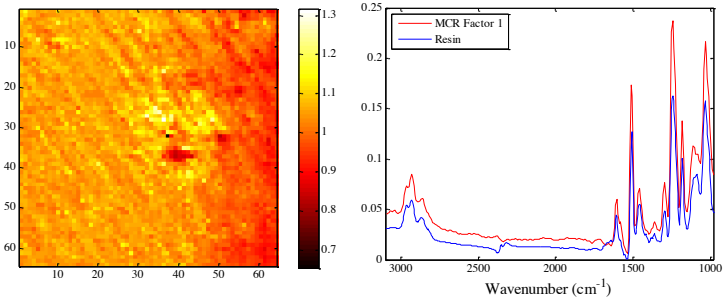
Example of Dealing w/ Clutter

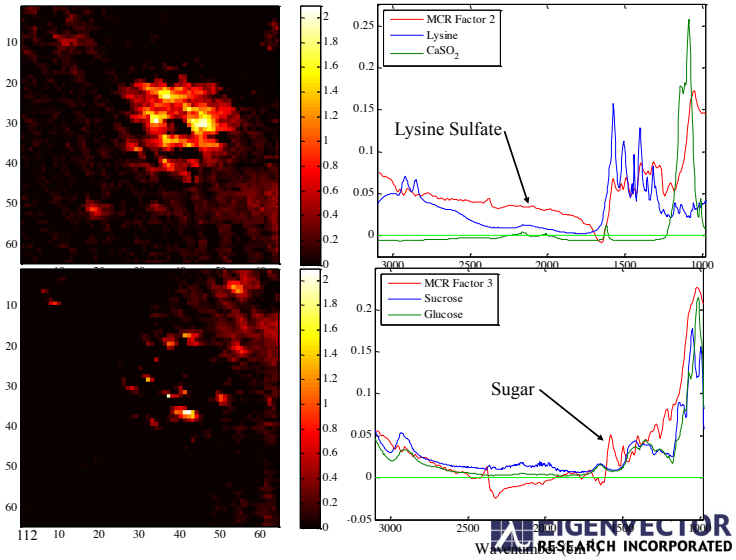
- MIA Example: Multivariate Curve Resolution (MCR)
  - Perform EMSC magnitude and slope correction (more later …)
    - reference is an estimate of the resin spectrum with robust fitting
    - allow glucose, lysine, $CaSO_4$ spectra to pass the filter
    - Gallagher, Blake, Gassman, *J. Chemometr.*, **19**(5-7), 271-281 (2005).
  - Step 2: Account for scratches using spatial constraints:
    - Scores from a PCA of region 2778 to 1790 cm^{-1} w/ 2nd derivative preprocessing capture variability due to scratch features
    - Equality constraints on **C**: components 4 to 11➔the scratches
      - Soft equality Constraints on **S**: components 1 to 3
        - » Factor 1: resin, Factor 2: lysine (w/~ $CaSO_4$), Factor 3: glucose
  - ➔linear mixture model referred to as an extended mixture model

$$\mathbf{X} = [\mathbf{C} \quad \mathbf{T}][\mathbf{S} \quad \mathbf{P}]^T + \mathbf{E}$$

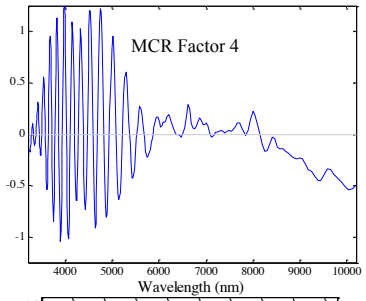
desired factors — interferences

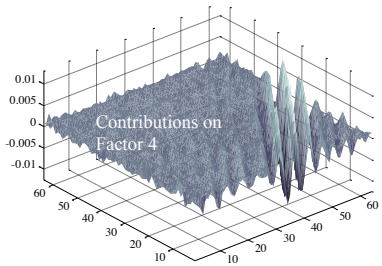
MCR Factor 1: Resin





Example of a Scratch Feature

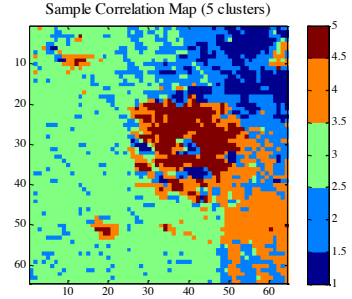




EIGENVECTOR RESEARCH INCORPORATED

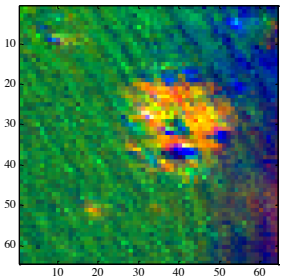
Contributions → RGB

C for Factors 1:3: 1-Norm Preprocessing
KNN Cluster Analysis of the MCR Contributions C



C for Factors [2 1 3] = RGB

R = lysine, G = resin, B = sucrose



Pei, L. Guilin, J., Davis, R.C., Shaver, J.M., Smentkowski, V.S., Asplund, M.C., Linford, M.R., *Applied Surface Science*, **253**(12), 5375-5386 (2007).

EIGENVECTOR RESEARCH INCORPORATED

ELS and EMSC Summary

- The extended mixture model can be used to account for complicated scatter
  - Best to have many analyte-free measurements to characterize clutter
- Weighted and Robust regression are useful for estimating EMSC model coefficients
- EISC and EMSC show promise
  - based in ILS and CLS respectively
  - spectra interpretable, relevant/predictive variance brought to top (scatter artifacts removed)

EIGENVECTOR RESEARCH INCORPORATED

External Parameter Orthogonaliziation

- EPO: form of the extended mixture model that can be used as a pre-processing by orthogonalizing to the clutter.
  - need a model of the clutter / interferences
    - e.g., spectra (as seen above) or PCA loadings

CLS model using the extended mixture model

desired factors interferences

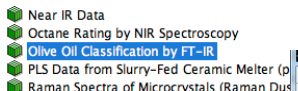
$$\mathbf{x} = \begin{bmatrix}\mathbf{c} & \mathbf{t}\end{bmatrix}\begin{bmatrix}\mathbf{S} & \mathbf{P}\end{bmatrix}^T + \mathbf{e}$$

$$\begin{bmatrix}\hat{\mathbf{c}} & \hat{\mathbf{t}}\end{bmatrix} = \mathbf{x}\begin{bmatrix}\mathbf{S} & \mathbf{P}\end{bmatrix}\left(\begin{bmatrix}\mathbf{S} & \mathbf{P}\end{bmatrix}^T\begin{bmatrix}\mathbf{S} & \mathbf{P}\end{bmatrix}\right)^{-1}$$

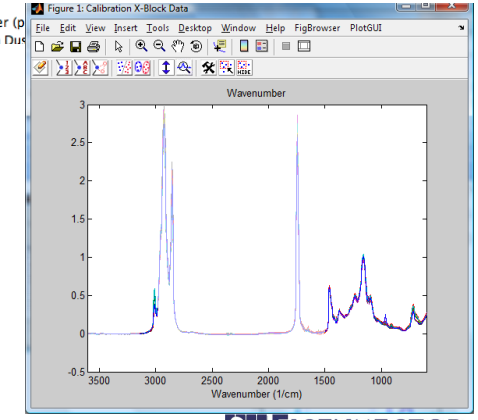
e.g., use a PCA model of intra-class variance to characterize the clutter

$$\mathbf{X}_{epo} = \mathbf{X}\left(\mathbf{I} - \mathbf{P}\mathbf{P}^T\right)$$

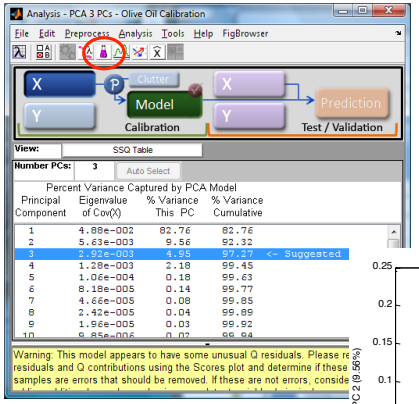
EIGENVECTOR RESEARCH INCORPORATED



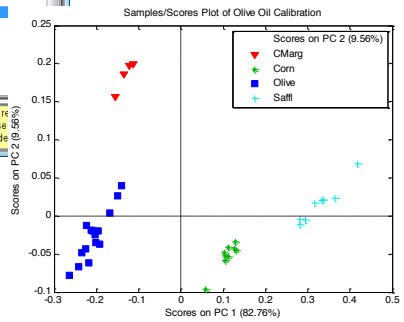
OliveOilData.mat
Four classes
Selected wavenumbers

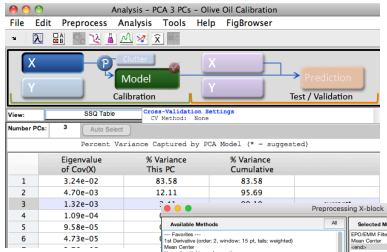


EIGENVECTOR RESEARCH INCORPORATED

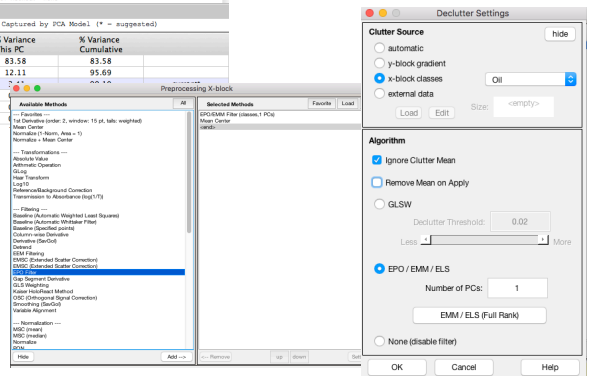


OliveOilData.mat
PCA of xcal
w/ mean-centering

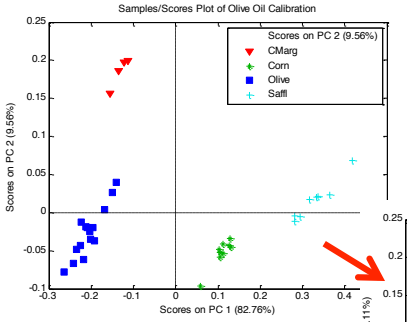




Set EPO to be based on 1 PC
of intra-class variance

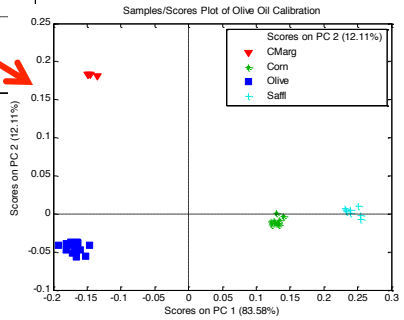


EIGENVECTOR RESEARCH INCORPORATED



EPO Results

OliveOilData.mat
PCA of xcal
w/ EPO and mean-centering

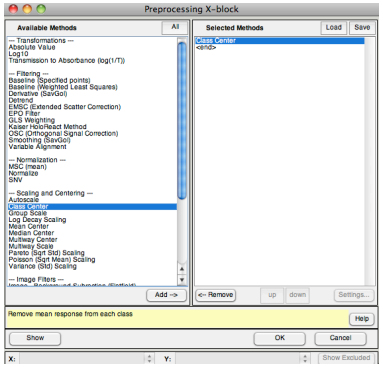


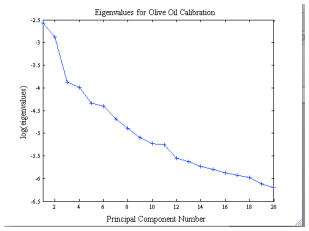
try it with the test data too….

EIGENVECTOR RESEARCH INCORPORATED

Why 1 PC for EPO Model?

How would the number of EPO PCs be determined?





EIGENVECTOR RESEARCH INCORPORATED

Preprocessing as Part of the Model

- Generally, preprocessing is applied prior to the modeling step (e.g., PCA, PLS)
- Some models include preprocessing as part of the model e.g.,
  - Maximum / Minimum Noise Fractions (MNF)
    - Green AA, Berman M, Switzer P, Craig MD (1988) IEEE Trans Geosci Remote Sens 26:65–74
  - Maximum Autocorrelation Factors (MAF)
    - same model as MNF, but clutter is defined differently
  - Generalized Least Squares (GLS)
    - Aitken, A., "On Least Squares and Linear Combinations of Observations", *Proceedings of the Royal Society of Edinburgh*, 1935, **55**, 42-48

EIGENVECTOR RESEARCH INCORPORATED

Generalized Least Squares (GLS)

- Generalized least squares (GLS) is a weighted version of the classical least squares (CLS) model

$$\mathbf{x}_{1\times N} = \mathbf{c}_{1\times K}\mathbf{S}^T_{K\times N} + \mathbf{e}_{1\times N}$$

CLS

$$\mathbf{e}\mathbf{e}^T = \left(\mathbf{x}-\mathbf{c}\mathbf{S}^T\right)\left(\mathbf{x}-\mathbf{c}\mathbf{S}^T\right)^T$$

$$\frac{d\left(\mathbf{e}\mathbf{e}^T\right)}{d\mathbf{c}} = -2\mathbf{S}^T\left(\mathbf{x}-\mathbf{c}\mathbf{S}^T\right) \Rightarrow 0$$

$$\mathbf{c} = \mathbf{x}\mathbf{S}\left(\mathbf{S}^T\mathbf{S}\right)^{-1}$$

GLS

$$\mathbf{e}\mathbf{W}_c^{-1}\mathbf{e}^T = \left(\mathbf{x}-\mathbf{c}\mathbf{S}^T\right)\mathbf{W}_c^{-1}\left(\mathbf{x}-\mathbf{c}\mathbf{S}^T\right)^T$$

$$\mathbf{c} = \mathbf{x}\mathbf{W}_c^{-1}\mathbf{S}\left(\mathbf{S}^T\mathbf{W}_c^{-1}\mathbf{S}\right)^{-1}$$

$$\mathbf{c} = \left(\mathbf{x}-\mathbf{1}\bar{\mathbf{x}}_c^T\right)\mathbf{W}_c^{-1}\mathbf{S}\left(\mathbf{S}^T\mathbf{W}_c^{-1}\mathbf{S}\right)^{-1}$$

$$\mathbf{W}_c = \frac{1}{M_c-1}\left(\mathbf{X}_c-\mathbf{1}\bar{\mathbf{x}}_c^T\right)^T\left(\mathbf{X}_c-\mathbf{1}\bar{\mathbf{x}}_c^T\right)$$

clutter covariance matrix

EIGENVECTOR RESEARCH INCORPORATED

CLS vs. GLS

- Comparison of CLS and GLS models

	$\mathbf{x}_w = \mathbf{x}\mathbf{W}_c^{-1/2}$
$\mathbf{X} = \mathbf{C}\mathbf{S}^T$	$\mathbf{S}_w = \mathbf{W}_c^{-1/2}\mathbf{S}$
$\hat{\mathbf{c}} = \mathbf{x}\mathbf{S}\left(\mathbf{S}^T\mathbf{S}\right)^{-1}$	$\hat{\mathbf{c}} = \mathbf{x}_w\mathbf{S}_w\left(\mathbf{S}_w^T\mathbf{S}_w\right)^{-1}$

Weighting by the inverse square root of the clutter covariance reduces the GLS model to CLS with weighted measurements and spectra i.e., it is a preprocessing step! The weighting can be viewed as a preprocessing step i.e., a pre-whitening.

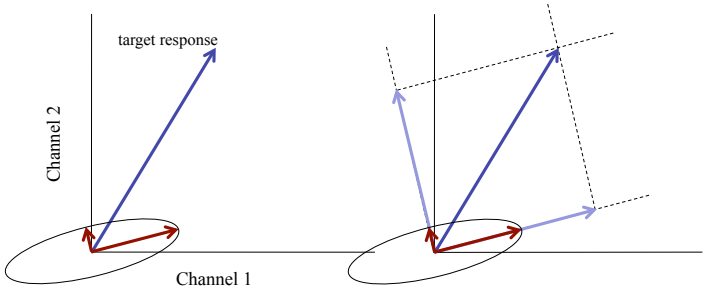
$\mathbf{X}_w = \mathbf{X}\mathbf{W}_c^{-1/2}$ can be used w/ PCA and ILS models (PLS, PCR)

N.B. Gallagher, "Detection, Classification and Quantification in Hyperspectral Images using Classical Least Squares Models," in *Techniques and Applications of Hyperspectral Image Analysis*, H. F. Grahn and P. Geladi, eds. (John Wiley & Sons, West Sussex, England), 181-201, 2007.

H. Martens, M. Høy, B.M. Wise, R. Bro and P.B. Brockhoff, "Pre-whitening of data by covariance-weighted pre-processing," *J. Chemo.*, **17**(3), 153-165 (2003).

EIGENVECTOR RESEARCH INCORPORATED

How the Whitening Works: Target Projected onto Clutter Directions

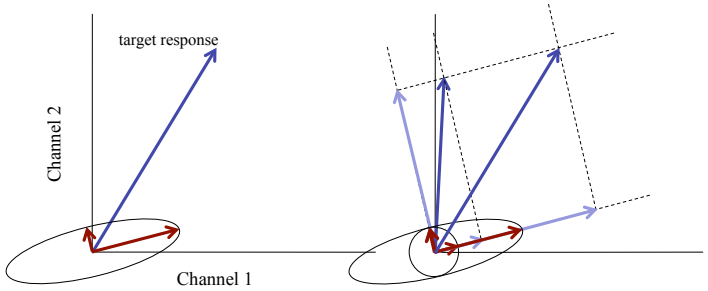


PCA of (correlated) clutter

125

EIGENVECTOR RESEARCH INCORPORATED

Scale Target by Clutter

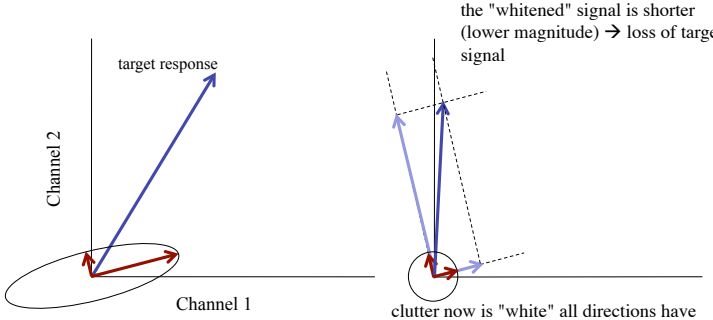


PCA of (correlated) clutter

126

EIGENVECTOR RESEARCH INCORPORATED

Whitened Signal



PCA of (correlated) clutter

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

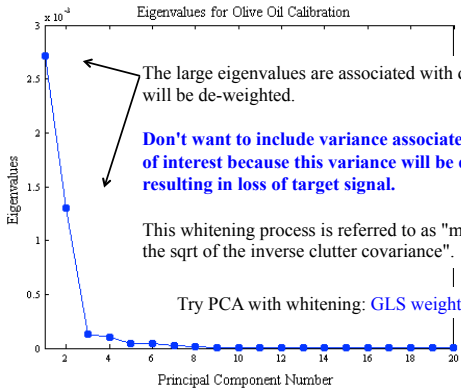
clutter now is "white" all directions have similar variance

127

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.

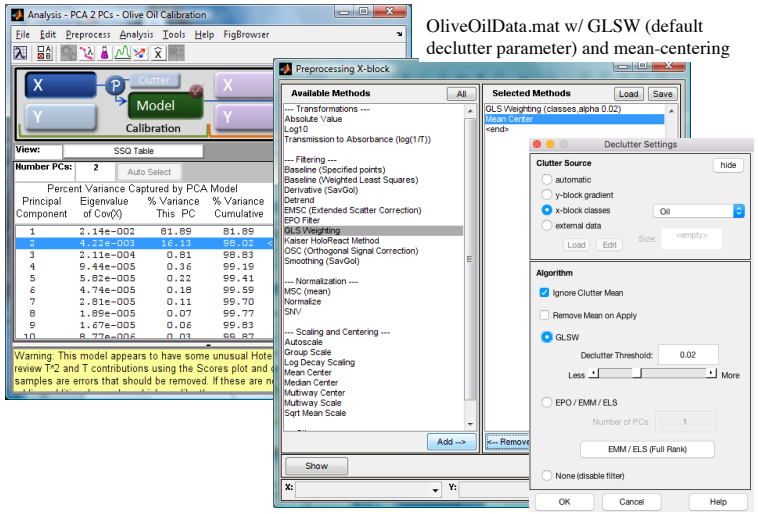
Don't want to include variance associated with target of interest because this variance will be de-weighted resulting in loss of target signal.

This whitening process is referred to as "multiplying by the sqrt of the inverse clutter covariance".

Try PCA with whitening: GLS weighting → GLSW

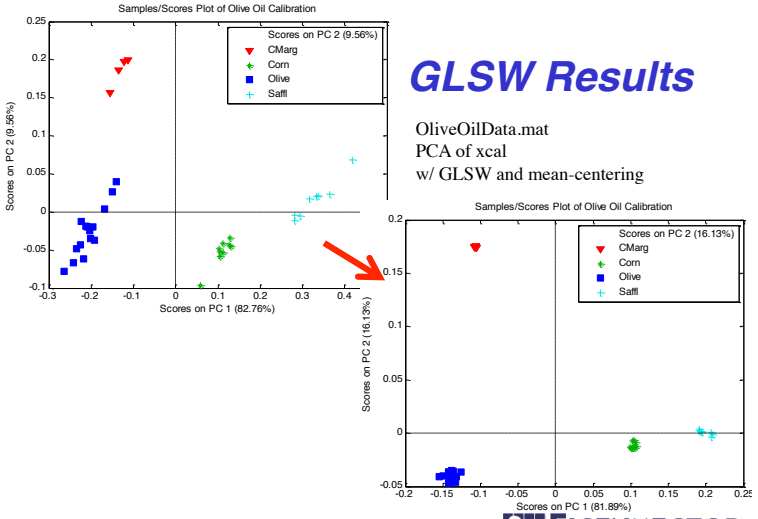
128

EIGENVECTOR RESEARCH INCORPORATED



OliveOilData.mat w/ GLSW (default declutter parameter) and mean-centering

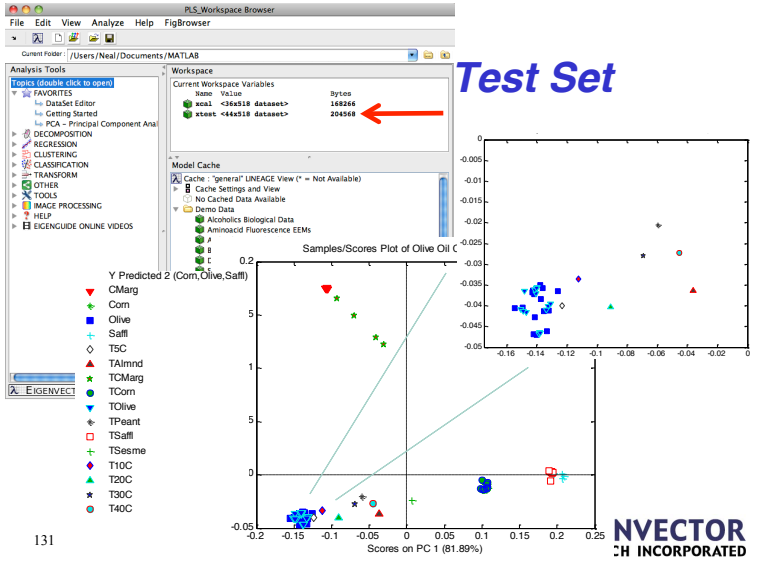
EIGENVECTOR RESEARCH INCORPORATED



GLSW Results

OliveOilData.mat
PCA of xcal
w/ GLSW and mean-centering

EIGENVECTOR RESEARCH INCORPORATED



Test Set

EIGENVECTOR RESEARCH INCORPORATED

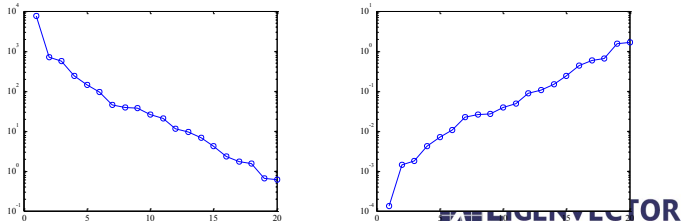
Inverse Clutter Covariance

- The SVD can be used to obtain the inverse of the clutter covariance
- Interpreting the inverse isn't necessarily intuitive

$$\mathbf{W}_c = \frac{1}{M_c - 1}\left(\mathbf{X}_c - \mathbf{1}\bar{\mathbf{x}}_c^T\right)^T\left(\mathbf{X}_c - \mathbf{1}\bar{\mathbf{x}}_c^T\right) \qquad \mathbf{W}_c^{-1} = \left(\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T\right)^{-1} = \left(\mathbf{V}^T\right)^{-1}\boldsymbol{\Lambda}^{-1}\mathbf{V}^{-1}$$

$$\mathbf{W}_c = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T \qquad = \mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^T$$

$$\mathbf{W}_c^{-\frac{1}{2}} = \mathbf{V}\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{V}^T$$



EIGENVECTOR RESEARCH INCORPORATED

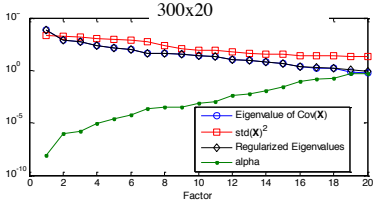
Regularization

- If an inverse doesn't exist or is ill-conditioned regularization is necessary.
  - ridging
  - **D** diagonal

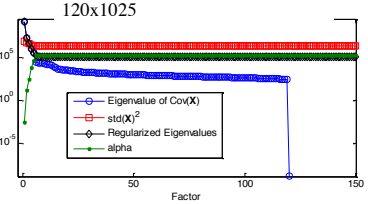
$$\mathbf{W}_c^{-1} = \mathbf{V}\left(\mathbf{\Lambda} + \alpha^2\mathbf{I}\right)^{-1}\mathbf{V}^T$$

$$\mathbf{W}_c^{-1} = \mathbf{V}\left(\mathbf{\Lambda} + \mathbf{D}\right)^{-1}\mathbf{V}^T$$

cov_cv(xblock1.data)
300x20

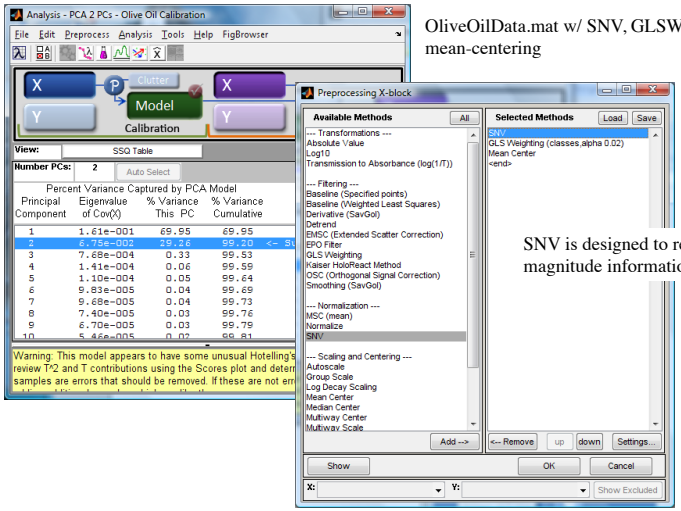


cov_cv(raman_dust_particles.data)
120x1025



EIGENVECTOR RESEARCH INCORPORATED

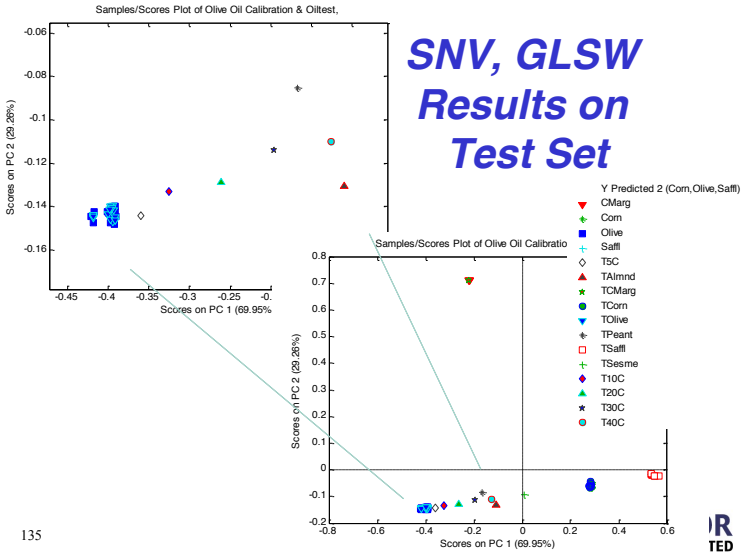
OliveOilData.mat w/ SNV, GLSW and mean-centering



SNV is designed to remove magnitude information

EIGENVECTOR RESEARCH INCORPORATED

SNV, GLSW Results on Test Set

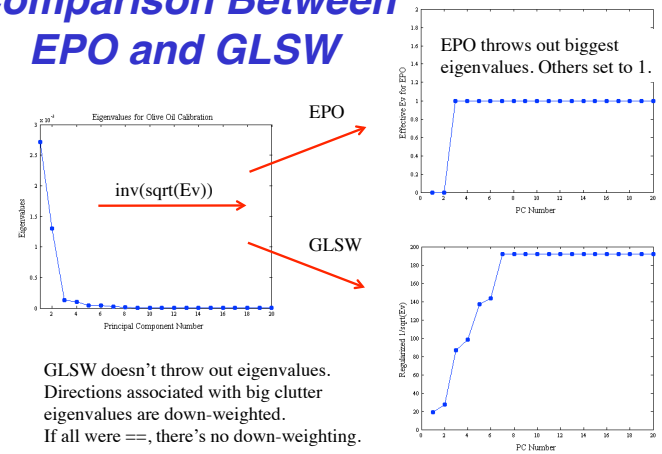


EPO & GLS Weighting Results

- EPO and GLS weighting can be viewed as complimentary approaches
  - both are designed to account for clutter / interferences
    - EPO is a strict orthogonalization - directions associated with clutter PCs are completely removed from the data
    - GLS de-weights the clutter directions
      - clutter PCs with very large eigenvalue compared to the rest are effectively removed
      - directions associated with medium eigenvalues remain but are de-weighted
  - neither is designed to account for magnitude differences that might make a class cluster spread out

EIGENVECTOR RESEARCH INCORPORATED

Comparison Between EPO and GLSW



GLSW doesn't throw out eigenvalues. Directions associated with big clutter eigenvalues are down-weighted. If all were ==, there's no down-weighting.

EIGENVECTOR RESEARCH INCORPORATED

Maximum Noise Fractions Minimum Noise Factors

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

$\mathbf{X}_{M_x xN}$ measured data with mean $\mathbf{x}_{1xN}$

$\mathbf{X}_{c,M_c xN}$ clutter matrix with mean $\bar{\mathbf{x}}_{c,Nx1}$

$\boldsymbol{\Sigma}_X = \frac{1}{(M_X - 1)}\left(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T\right)^T\left(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^T\right)$ signal covariance

$\boldsymbol{\Sigma}_C = \frac{1}{(M_C - 1)}\left(\mathbf{X}_c - \mathbf{1}\bar{\mathbf{x}}_c^T\right)^T\left(\mathbf{X}_c - \mathbf{1}\bar{\mathbf{x}}_c^T\right)$ clutter covariance

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

EIGENVECTOR RESEARCH INCORPORATED

MNF Derivation

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

The objective function is a scalar function.

Taking the derivative wrt $\mathbf{v}$, and setting to 0 gives:

$$\frac{\boldsymbol{\Sigma}_X \mathbf{v}_i\left(\mathbf{v}_i^T \boldsymbol{\Sigma}_C \mathbf{v}_i\right) - \boldsymbol{\Sigma}_C \mathbf{v}_i\left(\mathbf{v}_i^T \boldsymbol{\Sigma}_X \mathbf{v}_i\right)}{\left(\mathbf{v}_i^T \boldsymbol{\Sigma}_C \mathbf{v}_i\right)^2} = 0$$

Rearranging results in the MNF eigenvector solutions.

$$\boldsymbol{\Sigma}_X \mathbf{v}_i = \left(\frac{\mathbf{v}_i^T \boldsymbol{\Sigma}_X \mathbf{v}_i}{\mathbf{v}_i^T \boldsymbol{\Sigma}_C \mathbf{v}_i}\right)\boldsymbol{\Sigma}_C \mathbf{v}_i$$

$\boldsymbol{\Sigma}_X \mathbf{v}_i = \lambda_i \boldsymbol{\Sigma}_C \mathbf{v}_i$ generalized eigenvector problem

$\boldsymbol{\Sigma}_C^{-1} \boldsymbol{\Sigma}_X \mathbf{v}_i = \lambda_i \mathbf{v}_i$ eigenvector problem with non-symmetric matrix (there is a trick to convert to a symmetric eigenvector problem)

EIGENVECTOR RESEARCH INCORPORATED

MNF vs PCA

	PCA	MNF
objective function	$\max_{\mathbf{v}_i \neq 0}\left(\mathbf{v}_i^T \boldsymbol{\Sigma}_X \mathbf{v}_i\right)$ subject to $\mathbf{v}_i^T \mathbf{v}_i = 1$	$\max_{\mathbf{v}_i \neq 0}\left(\frac{\mathbf{v}_i^T \boldsymbol{\Sigma}_X \mathbf{v}_i}{\mathbf{v}_i^T \boldsymbol{\Sigma}_C \mathbf{v}_i}\right)$
eigen-problem	$\boldsymbol{\Sigma}_X \mathbf{v}_i = \lambda_i \mathbf{v}_i$	$\boldsymbol{\Sigma}_X \mathbf{v}_i = \lambda_i \boldsymbol{\Sigma}_C \mathbf{v}_i$
with clutter (GLSW)	$\boldsymbol{\Sigma}_C^{-1/2} \boldsymbol{\Sigma}_X \boldsymbol{\Sigma}_C^{-1/2} \mathbf{v}_i = \lambda_i \mathbf{v}_i$	$\boldsymbol{\Sigma}_C^{-1} \boldsymbol{\Sigma}_X \mathbf{v}_i = \lambda_i \mathbf{v}_i$

Gallagher, N.B., Shaver, J.M., Bishop, R., Roginski, R.T., Wise, B.M., "Decompositions with Maximum Signal Factors," *J. Chemometr.*, **28**(8), 663-671 (2014), DOI: 10.1002/cem.2634.

EIGENVECTOR RESEARCH INCORPORATED

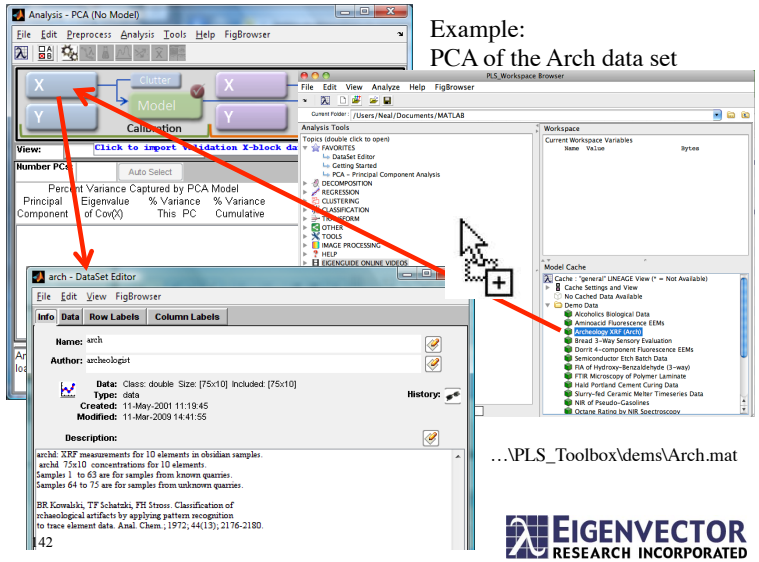
Example: ARCH

- 10 Variables: metal concentration (ppm via XRF)
- 75 Samples:
  - 63 obsidian samples from 4 quarries (known origin)
    - used to calibrate the models
    - center each cluster to own mean to estimate clutter
  - 12 artifacts (unknown origin)
    - which of the 4 known quarries do they belong?
- Data Matrix **X** is 75 by 10

141 arch.mat

EIGENVECTOR RESEARCH INCORPORATED

Example:
PCA of the Arch data set



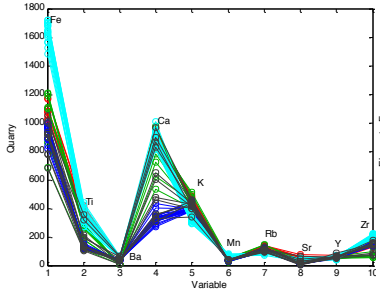
…\PLS_Toolbox\dems\Arch.mat

142

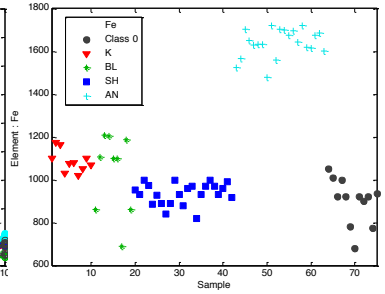
EIGENVECTOR RESEARCH INCORPORATED

Plot the Data

Plot data
Plot:Rows
View:Classes:Quarry
View:Labels:Element

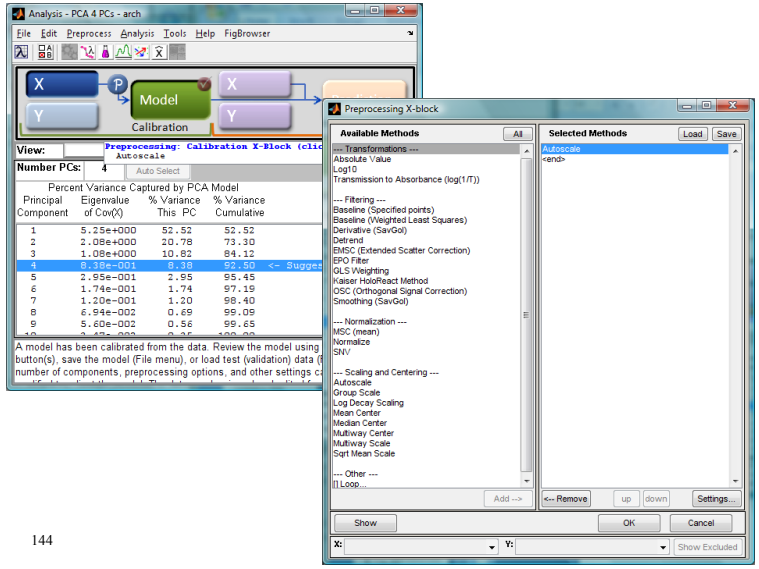


Plot data
Plot:Columns
View:Classes:Quarry
View:Labels:uncheck



143

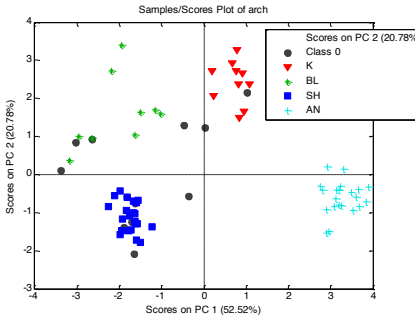
EIGENVECTOR RESEARCH INCORPORATED



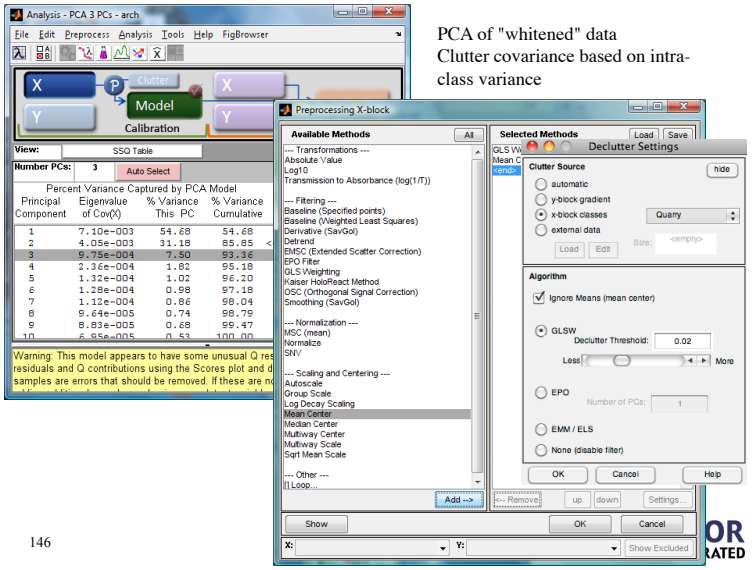
144

Plot the Scores

Plot scores
Select PC 2 vs PC 1
No Confidence Limit
Insert Legend
View:Spawn Static View

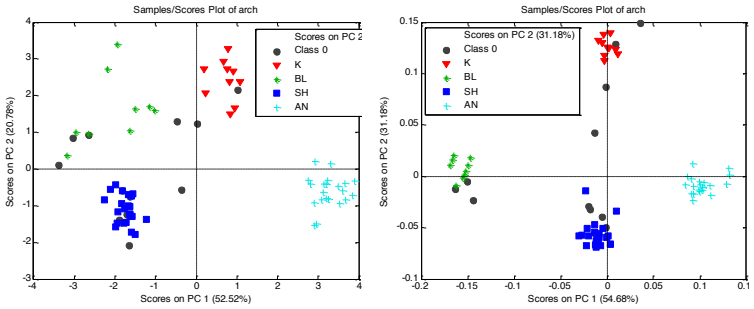


EIGENVECTOR RESEARCH INCORPORATED



PCA of "whitened" data
Clutter covariance based on intra-class variance

PCA vs PCA w/ Whitened Data

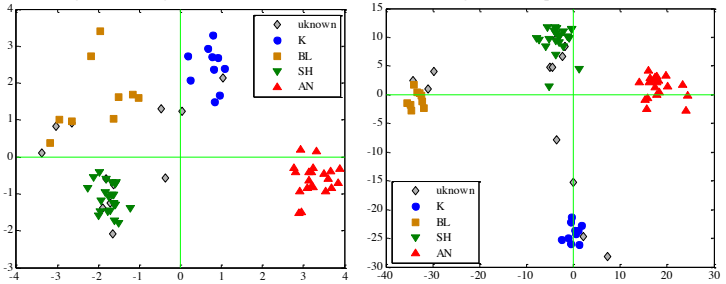


EIGENVECTOR RESEARCH INCORPORATED

PCA vs MNF Scores

PCA scores
scores for calibration set are orthogonal
loadings are orthogonal

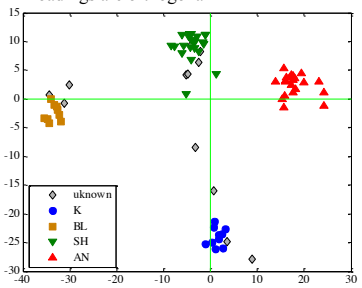
MNF scores
scores for calibration set are oblique
loadings are oblique



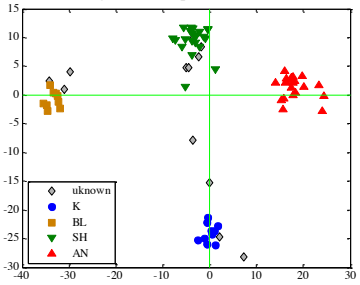
EIGENVECTOR RESEARCH INCORPORATED

PCA pre-whitened vs MNF

PCA scores after GLS weighting
scores for calibration set are orthogonal
loadings are orthogonal



MNF scores
scores for calibration set are oblique
loadings are oblique



149ArchMNFGLS.m

EIGENVECTOR RESEARCH INCORPORATED

Maximum Autocorrelation Factors for Multivariate Images

- For MNF, the clutter was intra-class variance
- For MAF, the clutter is the first spatial difference
  - the first difference should be high on edges and just noise w/in clusters
  - the result is the same generalized eigenvector problem as MNF with different clutter $\mathbf{\Sigma}_C$

T.A. Blake, J.F. Kelly, N.B. Gallagher, P.L. Gassman and T.J. Johnson, "Passive detection of solid explosives in Mid-IR hyperspectral images," *Anal Bioanal Chem*, **395**, 337-348, 2009.

N.B. Gallagher, J.F. Kelly, T.A. Blake, "Passive infrared hyperspectral imaging for standoff detection of tetryl explosive residue on a steel surface," Whispers 2010, June 14-16, Reykjavik, Iceland

150

EIGENVECTOR RESEARCH INCORPORATED

TOF-SIMS of Time Release Drug Delivery System

- Multi-layer drug beads serve as a controlled-release delivery system
- TOF-SIMS measurements of a drug bead embedded in epoxy and sliced to expose the bead cross-section
- Examine integrity of the layers and distribution of ingredients

Thanks to Physical Electronics and Anna Belu for the data.
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

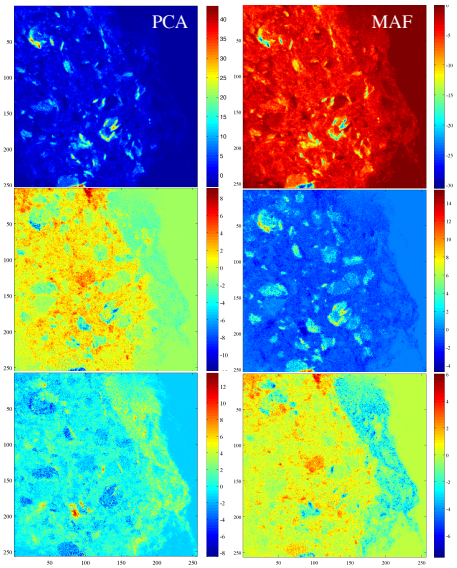
151

EIGENVECTOR RESEARCH INCORPORATED

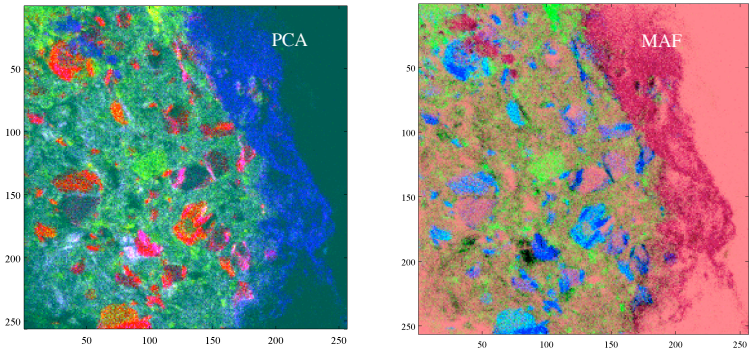
…\EVRIHW\Avicel.mat

PCA used Poisson scaling and mean centering.
MAF used no additional preprocessing.

Both models have a sign ambiguity:
$\mathbf{X} = \mathbf{TP}' = (-\mathbf{T})(-\mathbf{P}')$



152



RGB images after auto-contrasting

MAF can be applied to time-series as well.
Other models include principal autocorrelation factors, maximum difference factors.

EIGENVECTOR RESEARCH INCORPORATED

General Centering and Scaling

$\tilde{\mathbf{X}} = \mathbf{W}^{-1}\left(\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T\right)\boldsymbol{\Sigma}^{-1/2}$ general centering and scaling

$\mathbf{X}_{M_x xN}$ measured data $\quad \boldsymbol{\mu}_{Nx1}$ center

$\mathbf{W}_{M_x xM_x}$ sample weighting $\quad \boldsymbol{\Sigma}_{NxN}$ variable weighting

$\boldsymbol{\Sigma} = \mathbf{I} \quad \mathbf{W} = \mathbf{I} \quad \boldsymbol{\mu} = \mathbf{0}$ no scaling

$\boldsymbol{\Sigma} = \boldsymbol{\sigma}^2 \quad \mathbf{W} = \mathbf{I} \quad \boldsymbol{\mu} = \bar{\mathbf{x}}$ diagonal (autoscaling)

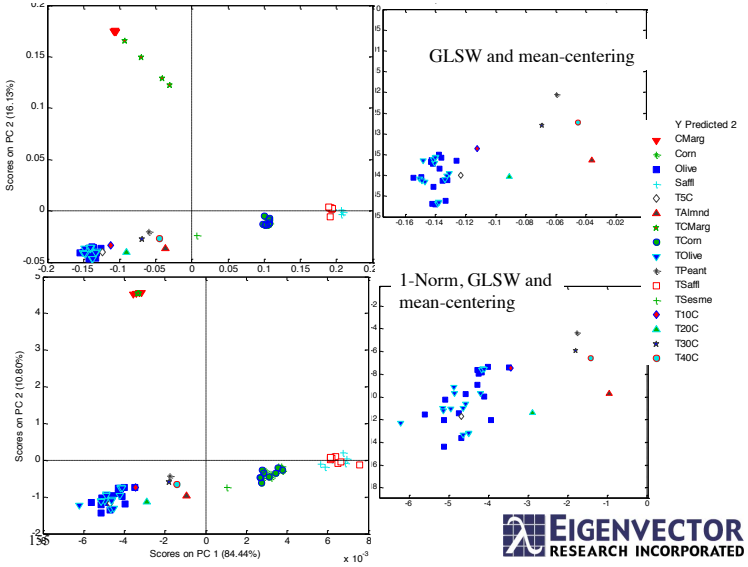
$\boldsymbol{\Sigma} = \mathbf{W}_c \quad \mathbf{W} = \mathbf{I} \quad \boldsymbol{\mu} = \bar{\mathbf{c}}$ GLS weighting

$\boldsymbol{\Sigma}_C = \frac{1}{(M_C - 1)}\left(\mathbf{C} - \mathbf{1}\bar{\mathbf{c}}\right)^T\left(\mathbf{C} - \mathbf{1}\bar{\mathbf{c}}\right)$ clutter covariance

$\mathbf{C}_{M_c xN}$ measured clutter matrix

$\bar{\mathbf{c}}_{1xN}$ clutter mean

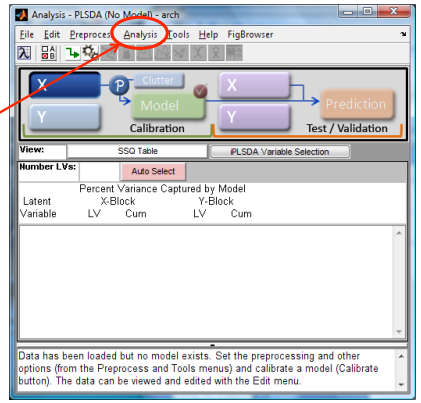
EIGENVECTOR RESEARCH INCORPORATED



EIGENVECTOR RESEARCH INCORPORATED

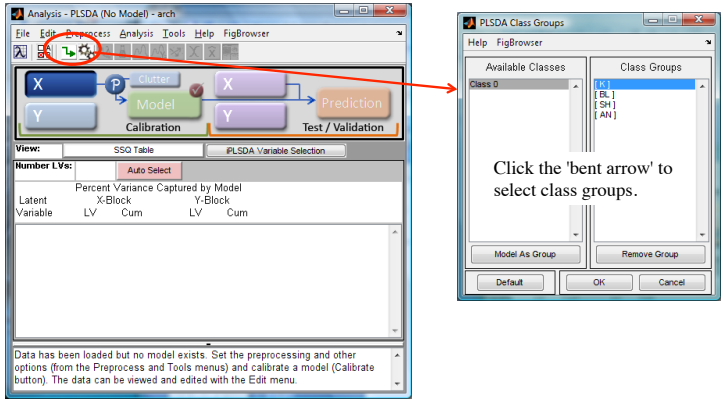
PLS-DA on ARCH

Under Analysis change the algorithm to PLSDA



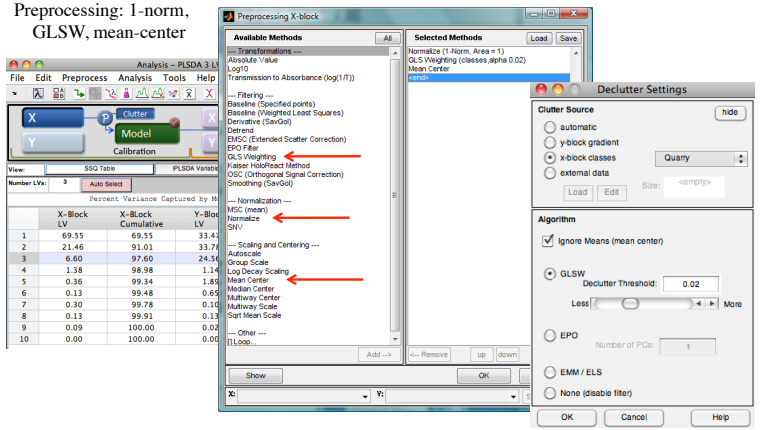
EIGENVECTOR RESEARCH INCORPORATED

PLS-DA on ARCH

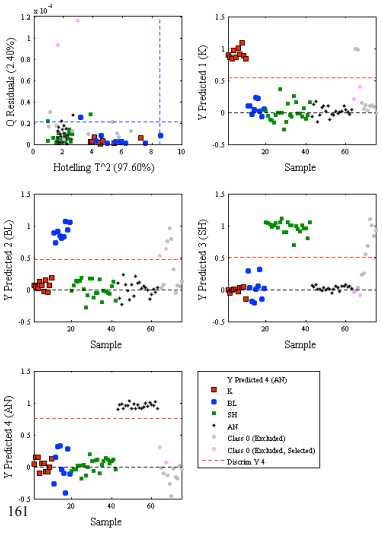


EIGENVECTOR RESEARCH INCORPORATED

Data: Arch
Analysis: PLSDA
Preprocessing: 1-norm, GLSW, mean-center



EIGENVECTOR RESEARCH INCORPORATED



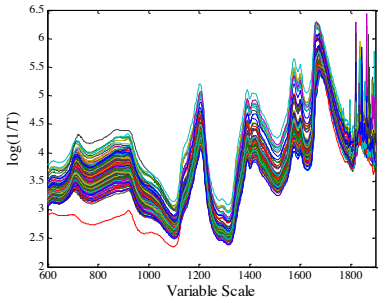
PLS-DA on Arch

- It's fairly easy to see which classes the samples might belong.
- Samples that don't belong to any class have high Q (selected points).

EIGENVECTOR RESEARCH INCORPORATED

NIR Shootout 2002

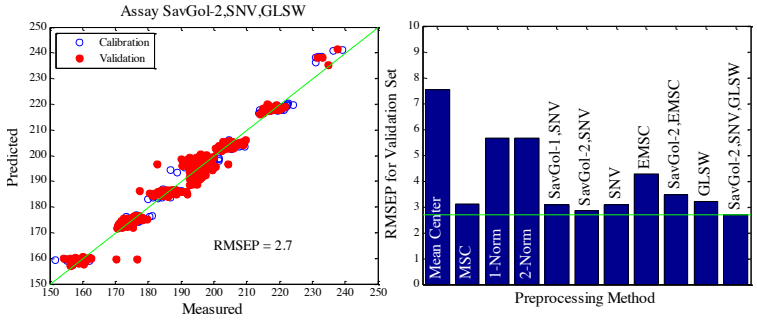
- Estimate assay value from NIR transmittance of tablets 600 to 1898 in 2 nm increments
  - http://www.idrc-chambersburg.org/shootout_2002.htm
  - Calibration (155 samples), Test (460 samples)



calibration data

EIGENVECTOR RESEARCH INCORPORATED

RMSEP on Validation Set



several of the simple approaches perform better than more complex pre-processings

EIGENVECTOR RESEARCH INCORPORATED

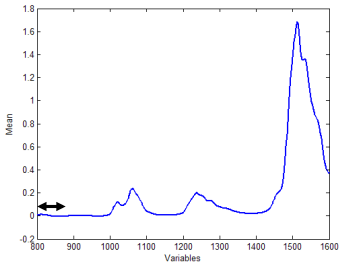
Summary up to now

- Centering and Offsets
  - choose appropriate "center" for the objective
- Normalization (removes magnitude)
- SNV, MSC (removes magnitude and offset)
- Baselining, Savitzky-Golay (high pass filter)
- EMSC (filter plus removes magnitude and offset)
- Autoscaling and Weighting
  - can be used to de-weight unusual variables, samples or directions
- ELS / EPO and GLS are complimentary procedures used to account for clutter
  - clutter needs to be characterized
  - not appropriate for everything - it does reduce net analyte signal
  - get rid instrument related problems first (may be simple) and then account for sampling related problems

EIGENVECTOR RESEARCH INCORPORATED

"Window" Methods - Filters

- Savitzky-Golay with derivative set to 0 is just a box-car average
- Box-car averages and other piece-wise (wavelength-localized) methods are filters
- Many preprocessing methods can also be done in a "windowed"/piecewise manner
  - OSC, MSC, continuos wavelets,….
- Many can be employed using convolution but some can not
  - median, max, min, …

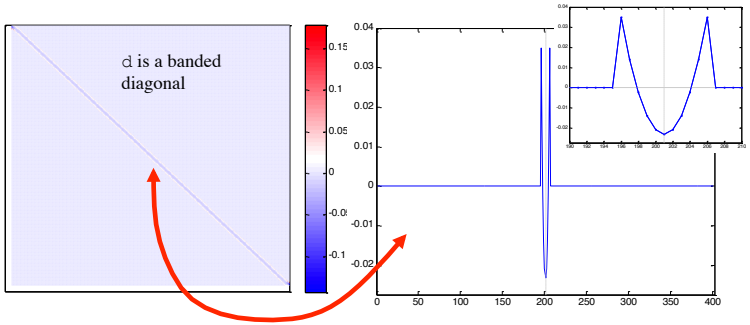


EIGENVECTOR RESEARCH INCORPORATED

SavGol as an Operator

```
>> x2 = spec1.data(1,:)+0.4+0.001*mncn(spec1.axisscale{2}')';
>> [d1,d] = savgol(x2,11,2,2);
>> pcolormap(fliplr(full(d)))
>> d_201 = full(d(:,201))';
```

d1 = x2*d;



EIGENVECTOR RESEARCH INCORPORATED

Derivatives via Convolution

- convolution
  - f is the spectrum
  - g is the point source function for the derivative

$$h(t) = \int_{-\infty}^{\infty} f(\tau) g(t-\tau) d\tau$$

$$H(\upsilon) = F(\upsilon) \cdot G(\upsilon)$$

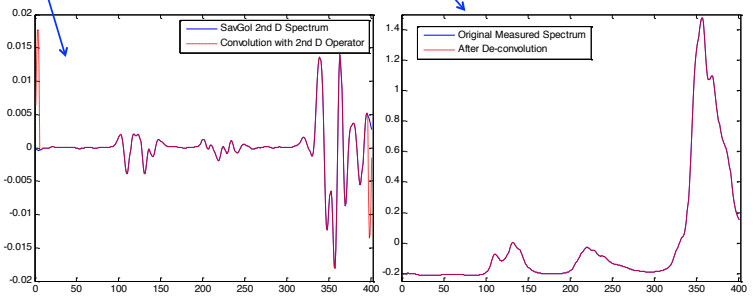
```
>> d = [d(201:end), d(2:201)];
```

  - h is the derivative
  - F, G, H are respective Fouier tranforms

EIGENVECTOR RESEARCH INCORPORATED

Compare SavGol and FFT

```
xf = line_filter(spec1(1,:),d);
xfu = line_filter(xf,d,struct('conv','deconvolve','reg',1e-11));
plot(spec1.data(1,:)-mean(spec1.data(1,:),2)), hold on
plot(xfu.data(1,:),'r--')
```



EIGENVECTOR RESEARCH INCORPORATED

Filters

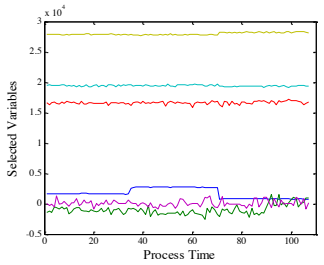
- Savitsky-Golay
  - For derivatives OR smoothing (noise reduction)
- Fourier
  - Remove high-frequency (noise) or low-frequency (baseline) components
  - Typically- NOT "windowed" but can be
    - Position (wavelength) information not considered
- Wavelets
  - Extracting information by BOTH frequency and position
    - Allows BOTH feature selection and pre-processing!
  - filters that are based on window-size (scale)
    - orthogonal and oblique basis functions can be used

LINE_FILTER

EIGENVECTOR RESEARCH INCORPORATED

Weakly Multi-Way Models

data from a single batch operation



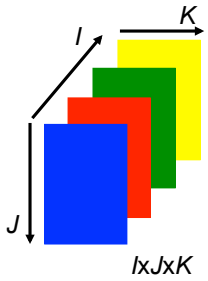
each process variable is a column of a data matrix

JxK

EIGENVECTOR RESEARCH INCORPORATED

Weakly Multi-Way Models

data from multiple batches are logically arranged into a cube



Batches x Time x Variable

data can be rearranged to two-way arrays in different ways (two are shown)

Batches*Time x Variable

arrangement for Multi-way PCA (MPCA)
each variable trajectory logically arranged in blocks

Batches x Variable*Time

EIGENVECTOR RESEARCH INCORPORATED

Weakly Multi-Way Models

centering: centers to the overall variable mean

autoscaling: scales to overall deviation about the overall mean

each column is a single variable

centering: centers to the variable mean trajectory (mean at any time)

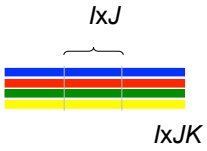
autoscaling: scales to deviation at each process time

each *I*x*J* block corresponds to a single variable's trajectory

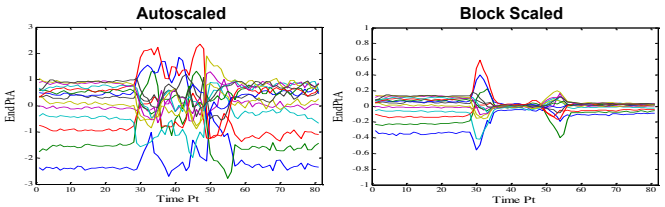
EIGENVECTOR RESEARCH INCORPORATED

Block / Group Scaling

- Autoscaling increases variance on irrelevant variables
  - e.g., during overetch (T>70) where little of interest occurs
- Scaling each of k=1,…,K blocks corresponding to a *I*x*J* set of trajectories for an individual variable
  - retains the relative variance within each variable's block
  - weighs each *I*x*J* variable block equally



Batches x Variable*Time



AUTO, GSCALE

EIGENVECTOR RESEARCH INCORPORATED

Block / Group Scaling

- With blocks of different variables, may want each block to have the same variance
  - Example: data set with NIR spectra and GC data and a collection of engineering variables, T, pH, P, Q etc.
- Variables within each block may be autoscaled or just mean-centered
- Determines a factor to multiply each block by so that total sum of squares (variance) is the same for each block (each variable)

see GSCALE

EIGENVECTOR RESEARCH INCORPORATED

Overall Summary 1/2

- Centering used when entire data set has an offset
  - Background subtraction is used when the offset varies sample-to-sample
- Column scaling can be used to weight variables
  - changes relative contribution to least-squares model
  - autoscaling (includes centering), Poisson scaling
- Filters attempt to remove (or partition) variance into different signals
  - want signal of interest to pass the filter
  - Savitzky-Golay, FFT, wavelets, …

EIGENVECTOR RESEARCH INCORPORATED

Overall Summary 2/2

- Modeling paradigm
  - design experiments to find signal *AND*
  - design experiment to characterize clutter
- When clutter can be characterized, then…
  - replicate samples with target not varying
  - off-target pixels in MIA
  - variables selective for clutter
- Models can explicitly account for clutter
  - Extended mixture model uses explicit interference factors
    - used in EMSC, ELS and EPO
  - Generalized least squares (GLS) weighted de-weights directions of high clutter (pre-whitening)
  - Generalized eigenvector problems (e.g., MAF, MNF)

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
- Always compare the effect of the pre-processing (RMSECV/RMSEP!) with the results from a model based on the raw data

EIGENVECTOR RESEARCH INCORPORATED

Pre-processing will offer…

- Models with better predictive performance (lower RMSEP) 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
  - The preprocessing itself can be overfit
    - Are the GLS weightings or interference factors relevant for future data? Use adaptive models for clutter?
- Pre-processing is seldom capable of saving a poor model but often makes a good model better

EIGENVECTOR RESEARCH INCORPORATED