CLS Regression Methods

(Building Interpretable Predictive Models)

EIGENVECTOR RESEARCH INCORPORATED

Outline

- Linear Mixture Model
- Classical Least Squares (CLS)
- Extended Least Squares (ELS)
- Weighted Least Squares (WLS)
- Generalized Least Squares (GLS)
- Constraints
- Misc.

EIGENVECTOR RESEARCH INCORPORATED

Course Materials

- These slides
- PLS_Toolbox or Solo 7.9 or later
- Data sets
  - From DEMS folder (installed with software)
    - Olive Oil Classification by FT-IR
      - Advanced Examples: plsdata (SFCM)
  - From EVRIHW folder (additional data sets)
    - EigenU_nir_data, SBRdata_EU

EIGENVECTOR RESEARCH INCORPORATED

Conventions & Notation

- *Rows* correspond to *samples*,
- *Columns* correspond to *variables*
- Notation:
  - $\mathbf{X}$ = matrix of predictor variables
  - $\mathbf{C}$ or $\mathbf{Y}$ = matrix of predicted variables
  - M = number of samples (observations)
  - N_x, N = number of $\mathbf{X}$ variables, K, N_c = number of $\mathbf{C}$ variables
  - $\mathbf{T}$ = $\mathbf{X}$-block scores matrix, $\mathbf{t}_1, \mathbf{t}_2, ..., \mathbf{t}_K$ score vectors
  - $\mathbf{P}$ = $\mathbf{X}$-block loads matrix, $\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_K$ loadings vectors
  - $\mathbf{S}$ = $\mathbf{X}$-block signal matrix, $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_K$ signal vectors
  - α = penalty parameter

EIGENVECTOR RESEARCH INCORPORATED

Linear Mixture Model

- **S** is a matrix corresponding to measurements to individual stimuli at unit response
  - spectra: multicomponent Beer's Law
    - from a library, estimated from the data (e.g., with MCR)
  - process response(s) obtained using DOE
  - linear mixture model
    - source apportionment
- **c** is a vector of coefficients
  - concentrations, contributions, coefficients, …

EIGENVECTOR RESEARCH INCORPORATED

Linear Mixture Model

$$\mathbf{x} = c_1\mathbf{s}_1 + c_2\mathbf{s}_2 + \ldots + c_K\mathbf{s}_K + \mathbf{e} = \begin{bmatrix} \mathbf{s}_1 & \mathbf{s}_2 & \ldots & \mathbf{s}_K \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_K \end{bmatrix} + \mathbf{e}$$

$$\mathbf{x} = \mathbf{Sc} + \mathbf{e}$$

$\mathbf{x}_{N\times 1} =$ measurement [column vector (it is a row of $\mathbf{X}_{M\times N}$)]

$\mathbf{c}_{K\times 1} =$ coefficients, contributions

$\mathbf{S}_{N\times K} =$ unit responses

$\mathbf{e}_{N\times 1} =$ residuals

$$\mathbf{X} = \mathbf{CS}^T + \mathbf{E}$$

$\mathbf{X}_{M\times N} =$ measurements

$\mathbf{C}_{M\times K} =$ coefficients

$\mathbf{S}_{N\times K} =$ unit responses

$\mathbf{E}_{M\times N} =$ residuals

EIGENVECTOR RESEARCH INCORPORATED

Advantages of Linear Mixture Model

- Interpretability
  - often the individual responses are interpretable
    - spectra or other physics
- Easy to incorporate prior information
  - useful constraints
    - e.g., non-negativity, closure, penalties, others …
- Model updating
  - can be fairly easy
    - interpretability helps here too

EIGENVECTOR RESEARCH INCORPORATED

PCA is a Linear Mixture Model

- PCA is a linear mixture model

$$\mathbf{X} = \mathbf{TP}^T + \mathbf{E}$$

$\mathbf{X}_{M\times N} =$ measurements

$\mathbf{T}_{M\times K} =$ scores

$\mathbf{P}_{N\times K} =$ loadings

$\mathbf{E}_{M\times N} =$ residuals

for the calibration data...

$\mathbf{T},\mathbf{P}$ are orthogonal

$\mathbf{C},\mathbf{S}$ are generally oblique

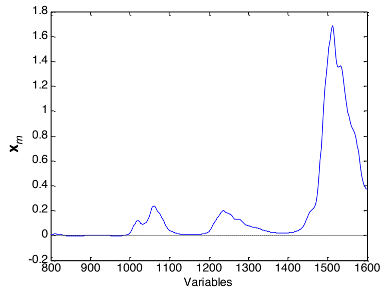
In PCA, the scores and loadings are calculated to maximize capture of variance **X** not to make predictions for **C**.

One way to obtain **C** and **S** is to use classical least squares (CLS).

EIGENVECTOR RESEARCH INCORPORATED

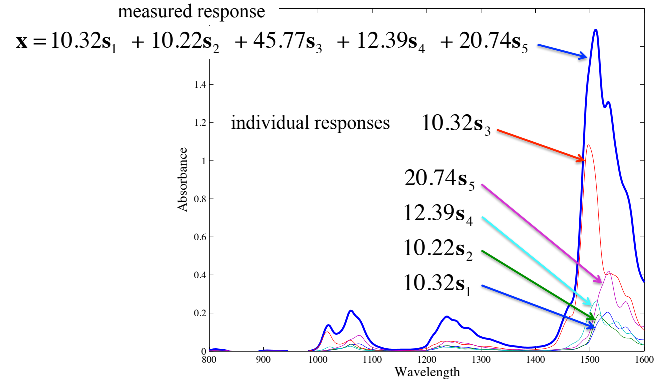
The CLS Model

- Given known pure component spectra, how much of each does it take to make up the observed m^{th} spectrum?
  - $\mathbf{x}_m = \mathbf{c}_m\mathbf{S}^T + \mathbf{e}_m$
  - $m = 1,\ldots,M$
  - $\mathbf{c}_m=[c_{m,1}, c_{m,2}, \ldots, c_{m,K}]$
  - $k = 1,\ldots,K$



9

EIGENVECTOR RESEARCH INCORPORATED



10 Linear Mixture Model

EIGENVECTOR RESEARCH INCORPORATED

CLS (cont.)

- Once $\mathbf{S}$ (the spectral "basis") is known, $\mathbf{c}$, the degree to which each component contributes to a new sample $\mathbf{x}$, can be determined from

$$\mathbf{c} = \mathbf{x}\mathbf{S}^+$$

where $\mathbf{S}^+$ is the pseudo-inverse of $\mathbf{S}$, defined in CLS as

$$\mathbf{S}^+ = \mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}$$

- Problem: How to get $\mathbf{S}$?
  - library, estimate from calibration measurements

11

EIGENVECTOR RESEARCH INCORPORATED

Classical Least Squares

$$\mathbf{X} = \mathbf{C}\mathbf{S}^T + \mathbf{E}$$
$$\mathbf{X} = \mathbf{C}\mathbf{S}^T$$
$$\mathbf{X}\mathbf{S} = \mathbf{C}\mathbf{S}^T\mathbf{S}$$
$$\mathbf{X}\mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1} = \mathbf{C}$$
$$\mathbf{S}^+ = \mathbf{S}(\mathbf{S}^T\mathbf{S})^{-1}$$

- Note that $\mathbf{S}^T\mathbf{S}$ is KxK (analytes by analytes) and square

12

EIGENVECTOR RESEARCH INCORPORATED

Estimating S

- Sometimes, **S** can be compiled *a priori* from a data base/spectral library, or from direct measurements of pure components
  - Problem: must account for all components that can contribute to **X**!
- **S** can also be estimated from mixtures, provided all **C** are known and enough samples are available:

$$\mathbf{S}^T = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{X}$$

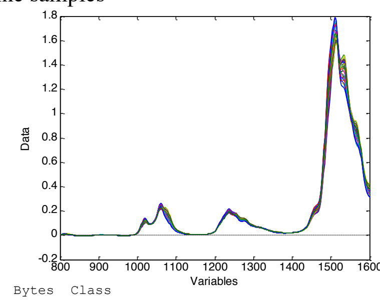
  - Problem: The concentration of *every analyte that contributes to* **X** must be known!*

*Interferences and unknowns can be handled with GLS or ELS type models, but their basis must be estimated.

13

CLS Example

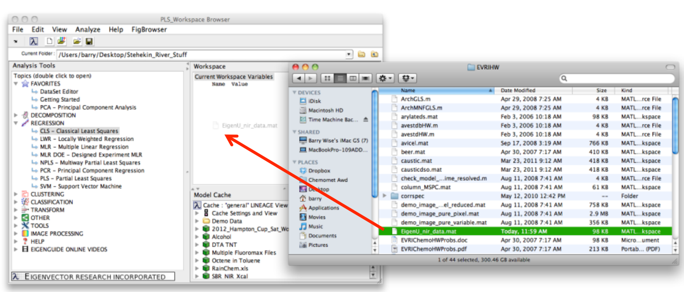
- NIR data of pseudo-gasoline samples
  - absorbance at 401 channels
  - 30 samples
  - 5 analytes
- EigenU_nir_data.mat
- Data broken into
  - 25 calibration samples and
  - 5 test samples




```
>> load EigenU_nir_data
>> whos
  Name            Size           Bytes  Class
  cal_conc        25x5           11002  dataset
  cal_spec        25x401         96466  dataset
  test_conc        5x5           10042  dataset
  test_spec        5x401         32146  dataset
```

14

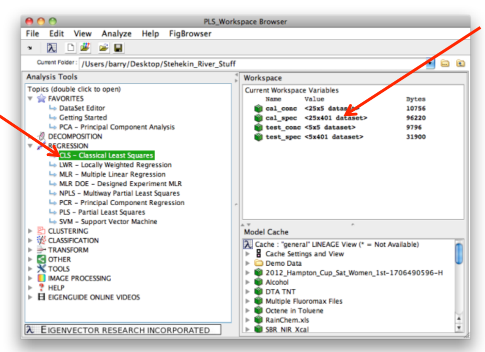
Load Data Into Browser



15

Start CLS Interface

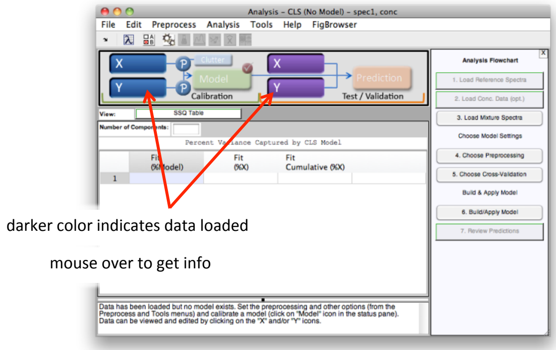
double-click



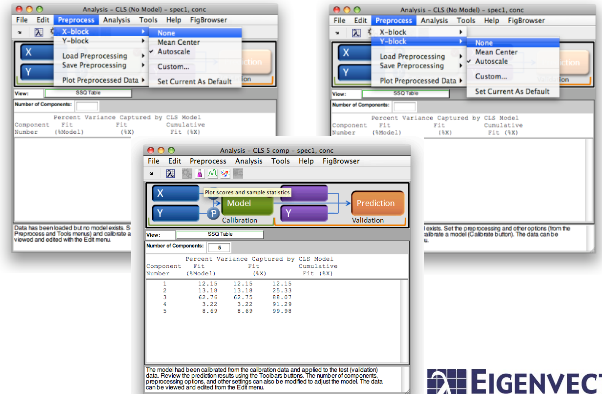
data in workspace

16

Data Loaded

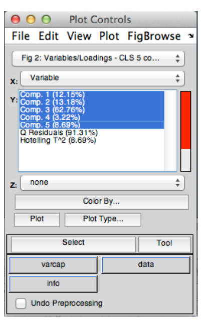


Set Preprocessing to "none," calculate model

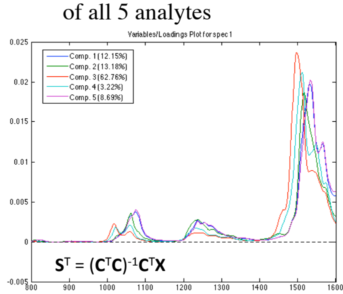


Pure Component Spectra

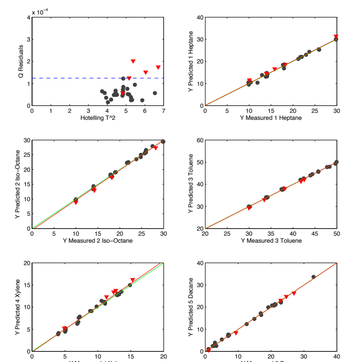
Click loadings "spectrum" icon, select all 5 components



S, estimated from mixtures, using known concentrations of all 5 analytes



Fit to Calibration and Estimate for Validation Samples



Click scores "flask" icon to get fits and predictions (test set).

Check "Show Cal Data with Test".

Calibration data (**black**)

Predicted test (**red**).

All analytes fit and predicted well.

Model Performance Measures

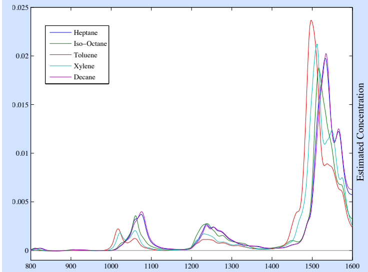
- Recall that root-mean square error is a measure of model performance
- Calibration
  - k is analyte index
- Prediction
  - test
- Cross-Validation
  - for J subsets

$$\mathrm{RMSEC}_k = \left\{ \frac{1}{M_{cal}} \sum_{m=1}^{M_{cal}} \left(c_{m,k} - \hat{c}_{m,k} \right)^2 \right\}^{1/2}$$

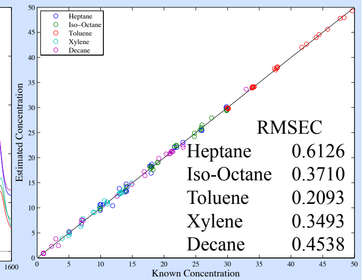
$$\mathrm{RMSEP}_k = \left\{ \frac{1}{M_{tst}} \sum_{m=1}^{M_{tst}} \left(c_{m,k} - \hat{c}_{m,k} \right)^2 \right\}^{1/2}$$

$$\mathrm{RMSECV}_k = \left\{ \sum_{j=1}^{J} \left(\frac{1}{M_{cal,j}} \sum_{m=1}^{M_{cal,j}} \left(c_{m,k} - \hat{c}_{m,k} \right)^2 \right) \right\}^{1/2}$$

Calibration: Pseudo-Gasoline



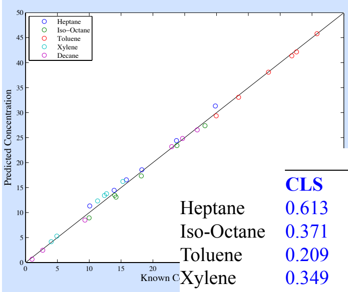
the estimated pure spectra are highly overlapped



	RMSEC
Heptane	0.6126
Iso-Octane	0.3710
Toluene	0.2093
Xylene	0.3493
Decane	0.4538

fits for the calibration data
RMSEC = root mean squared error of calibration

Prediction: Pseudo-Gasoline



LOO cross-validation to select LVs in PLS.

Not surprisingly, results between methods are similar – data are from a good DOE.

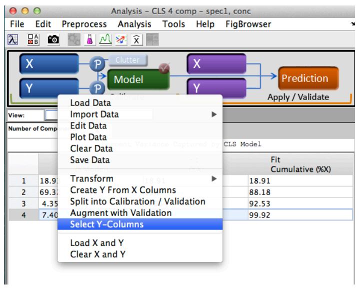
	RMSEC			RMSEP		
	CLS	**PLS2**	**PLS1**	**CLS**	**PLS2**	**PLS1**
Heptane	0.613	0.612	0.601	0.935	0.923	0.891
Iso-Octane	0.371	0.369	0.339	0.826	0.840	0.764
Toluene	0.209	0.208	0.225	0.392	0.399	0.430
Xylene	0.349	0.347	0.382	0.843	0.866	0.904
Decane	0.454	0.452	0.454	0.434	0.432	0.406

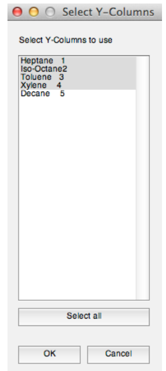
PLS2 used 5 LVs. PLS1 used 5 LVs except for Xylene that used 4 (as suggested by cross-validation, although 5 LVs worked better)

CLS Problem

- What if the concentration of 1 analyte was unknown?
- Repeat the CLS procedure using only the first 4 (of 5) analytes
- Attempt to predict concentrations of unused (test) samples

Select only the first four analytes and repeat

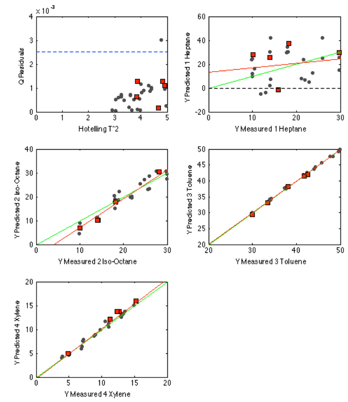




click 'cal Y: select Y-columns'

25

CLS Solution with One Analyte "Missing"

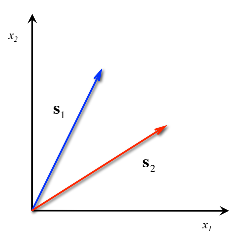


Click scores "flask" icon to get fits

Some analytes not fit (**black**) and not predicted (**red**) well, especially heptane

26

Spectra in "Two-Space"

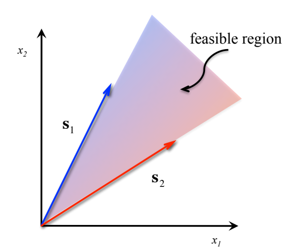


For two analytes measured at two spectral channels x_1 and x_2, the pure component spectra can be represented by $\mathbf{s}_1$ and $\mathbf{s}_2$.

All measurements in this two-space can be represented as a linear combination of $\mathbf{s}_1$ and $\mathbf{s}_2$.

27

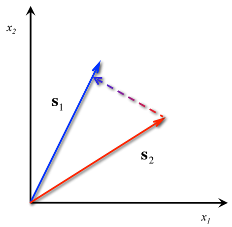
Non-negativity in "Two-Space"



Non-negativity forces all measurements to lie between $\mathbf{s}_1$ and $\mathbf{s}_2$.

28

CLS Problem in "Two-Space"



If the model only uses $\mathbf{s}_1$ then contribution to the signal from $\mathbf{s}_2$. will have a projection onto $\mathbf{s}_1$ resulting in poor predictions.

EIGENVECTOR RESEARCH INCORPORATED

The 'Problem' with CLS

- The "concentration of all the chromophores" must be known to account for them.
  - What to do? Is all lost?
  - For ILS we say, "the concentrations need not all be known but must vary if the model is to be robust to them."
    - This is the same for CLS
  - Implications for design of experiments …
    - vary both the analyte of interest *and* the interferences
      – useful for both ILS and CLS
    - can outside information be used? (e.g., pure spectra from a library)

EIGENVECTOR RESEARCH INCORPORATED

ALS for MCR (an aside)

- The alternating least-squares algorithm is one of the most popular for multivariate curve resolution.

given an initial guess $\mathbf{C}_0$

for $i = 1 : i_{\max}$

$$\mathbf{S}_i^T = \left(\mathbf{C}_{i-1}^T \mathbf{C}_{i-1}\right)^{-1} \mathbf{C}_{i-1}^T \mathbf{X}$$

$$\mathbf{C}_i = \left(\mathbf{S}_i^T \mathbf{S}_i\right)^{-1} \mathbf{S}_i^T \mathbf{X}$$

end

often subjected to non-negativity constraints and normalization of the columns of $\mathbf{S}$

EIGENVECTOR RESEARCH INCORPORATED

Extended Mixture Model

- The extended mixture model models the interferences and the target analyte separately in a CLS model
  - extended least squares, ELS
  - in the spirit of "vary both the analyte of interest and the interferences"
- The interference "spectra" aren't always used explicitly, however a basis that spans the interference variation is used.

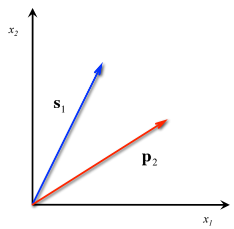
known spectra

spectra or basis that captures interference signal

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

EIGENVECTOR RESEARCH INCORPORATED

ELS / EMM in "Two-Space"



The interference model can be a pure component spectrum $\mathbf{s}_2$ or a PC $\mathbf{p}_2$.

P is intended to span the space of interferences, and be linearly independent of the known spectra **S**. Therefore **P** need not be PCs or spectra - these just tend to be convenient ways to capture interferences.

EIGENVECTOR RESEARCH INCORPORATED

Model of Interferences

- Assume that measurements can be made so that the target analyte contribution to the signal does not vary.
- The measured differences/variance is then due to interferences.
- Clutter = Interferences + noise
  - Clutter is all measured signal not related to the target of interest.

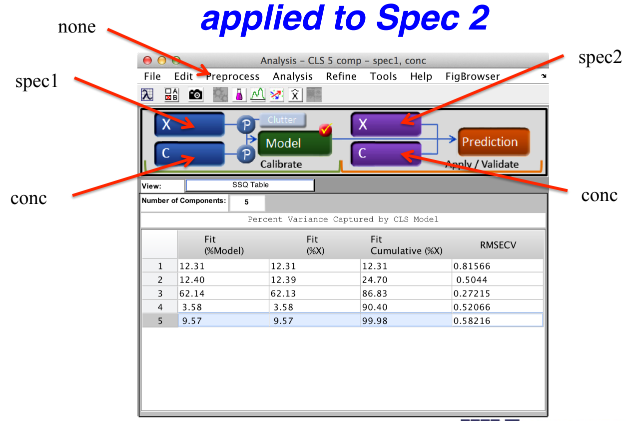
EIGENVECTOR RESEARCH INCORPORATED

ELS Example

- Build model on "spec1" from NIR pseudo-gasoline data
- Predict from "spec2"
  - Note that these are the same 30 samples measured on two different instruments
  - Data set used for standardization method tests

EIGENVECTOR RESEARCH INCORPORATED

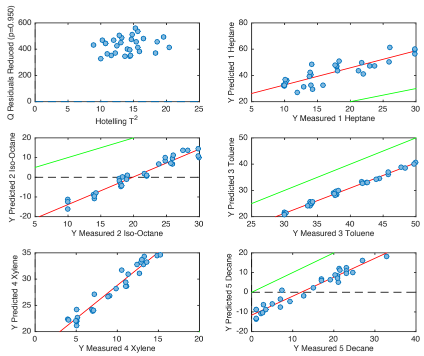
CLS Model of Spec 1 applied to Spec 2



	Fit (%Model)	Fit (%X)	Fit Cumulative (%X)	RMSECV
1	12.31	12.31	12.31	0.81566
2	12.40	12.39	24.70	0.5044
3	62.14	62.13	86.83	0.27215
4	3.58	3.58	90.40	0.52066
5	9.57	9.57	99.98	0.58216

EIGENVECTOR RESEARCH INCORPORATED

Results: not good!



EIGENVECTOR RESEARCH INCORPORATED

What's the Problem?

- Measured spectra looks different on second instrument compared to first
- This difference can be considered "clutter"
- Need to get a model of clutter
  - Mean difference
  - PCA basis of remaining differences

EIGENVECTOR RESEARCH INCORPORATED

Model of clutter

```
>> mean_dif = mean(spec1.data)-mean(spec2.data);
>> dif = mncn(spec1.data)-mncn(spec2.data);
>> [u,s,v] = svd(dif);
>> clutter_basis = [mean_dif; v(:,1:2)'];
```

EIGENVECTOR RESEARCH INCORPORATED

Extended Least Squares

- ELS using a clutter basis
  - Use PCA to get basis for clutter, **P**
  - **P** can be any basis with linearly independent columns
    - MCR could be used to obtain an interpretable basis

known spectra

spectra or basis that captures interference signal

define

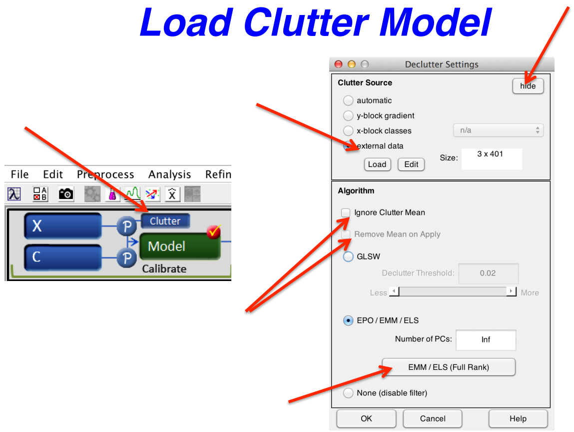
$$\begin{bmatrix} \mathbf{c}; & \mathbf{t} \end{bmatrix} \equiv \begin{bmatrix} \mathbf{c} \\ \mathbf{t} \end{bmatrix}$$

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

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

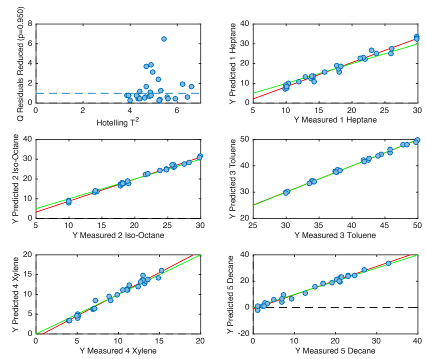
EIGENVECTOR RESEARCH INCORPORATED

Load Clutter Model



EIGENVECTOR RESEARCH INCORPORATED

Results with ELS: Much Better!



EIGENVECTOR RESEARCH INCORPORATED

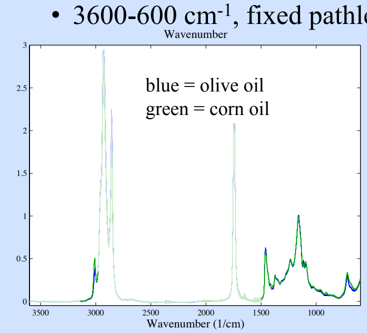
ELS Results

- Clutter basis allowed CLS model to account for the differences between original spectra (spec1) and new spectra (spec2)

EIGENVECTOR RESEARCH INCORPORATED

Olive Oil Example

- FTIR spectra of Olive and Corn oil
  - 3600-600 cm^{-1}, fixed pathlength NaCl



Can adulteration of olive oil with corn oil be detected?

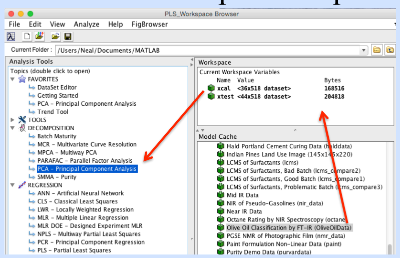
DB Dahlberg, SM Lee, SJ Wenger, JA Vargo, "Classification of Vegetable Oils by FT-IR," *Appl. Spectrosc.*, **51**(8), 1118-1124 (1997)

 Extended Least Squares-Olive Oil

EIGENVECTOR RESEARCH INCORPORATED

Olive Oil Example Details

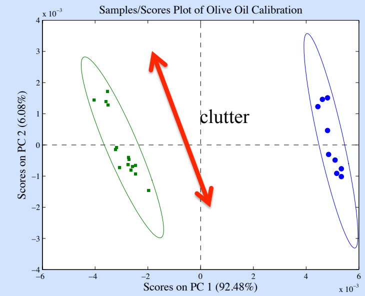
- load data into workspace
  - Olive Oil Classification by FT-IR (OliveOilData)
- drag xcal to PCA and plot the spectra



EIGENVECTOR RESEARCH INCORPORATED

PCA of Olive and Corn Oil Spectra

- 1-norm, Mean-center
  - 1-PC = 92%
- PCA can separate the pure oils
  - can it detect at low levels of corn oil?
  - the clutter looks correlated

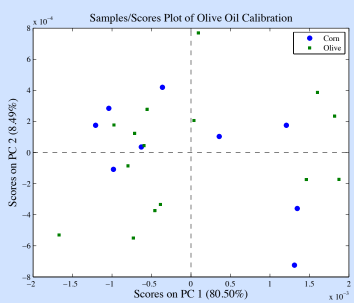


EIGENVECTOR RESEARCH INCORPORATED

PCA of Olive and Corn Oil Spectra

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

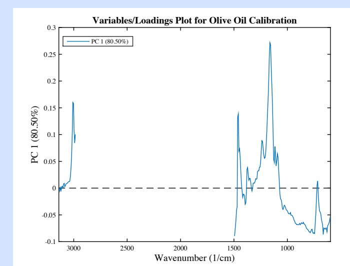
- 1-Norm and Class-Center
  - centers each class to its ***own*** mean
- Result is that the model is focused on intra-class variance
  - this is the clutter for this example
  - how many PCs to model the clutter?
  - exploratory analysis of the clutter may lead to insights for problem of interest

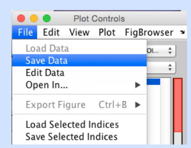


EIGENVECTOR RESEARCH INCORPORATED

PCA of Olive and Corn Oil Spectra

- 1 PC ~ 80%
- One factor captures a large fraction of the clutter variance.
  - PCs are orthogonal
  - true clutter "spectra" are unknown
- Save the loads

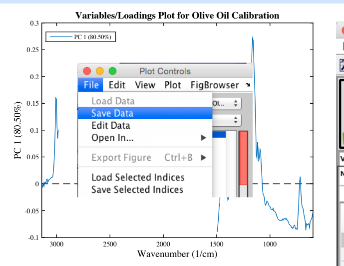


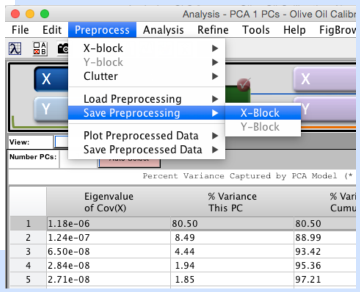


EIGENVECTOR RESEARCH INCORPORATED

PCA of Olive and Corn Oil Spectra

- Save the loads
- Save the preprocessed data





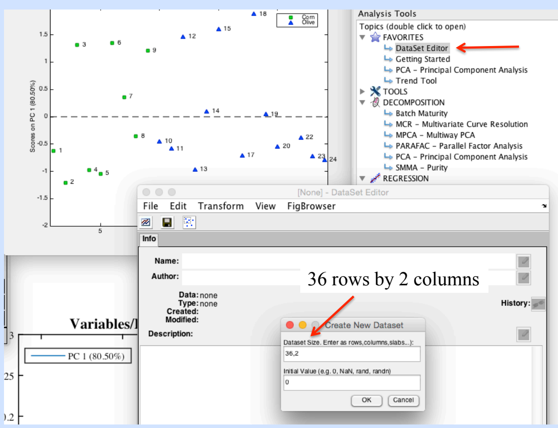
These saved files will be used later …

EIGENVECTOR RESEARCH INCORPORATED

CLS For Discrimination of the Oils

- The objective for this example is discrimination of the oils. Start by constructing a CLS model for each class.
- The model needs "concentrations"
  - will create a variable ycal that has 1's and 0's indicating "present" or "not present"

EIGENVECTOR RESEARCH INCORPORATED

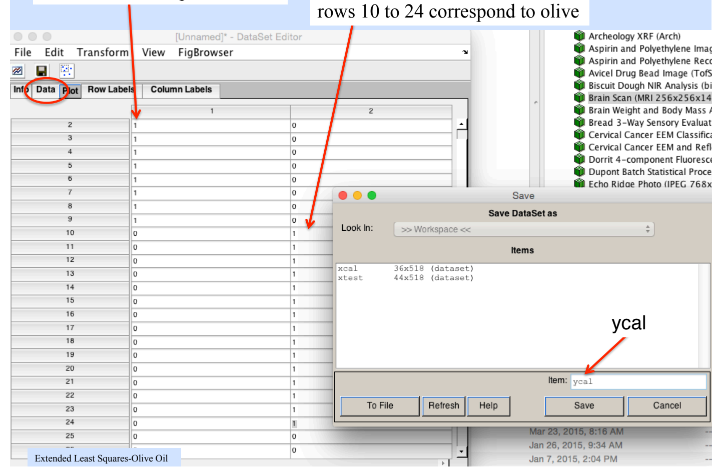


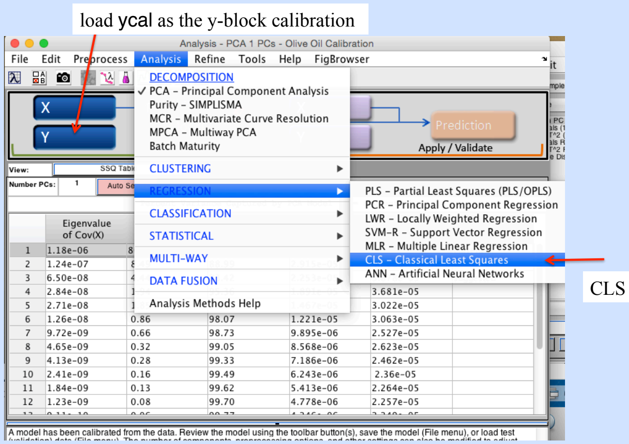
recall that xcal has 36 rows even though only the first 24 are being used (corn and olive oil)

EIGENVECTOR RESEARCH INCORPORATED

rows 1 to 9 correspond to corn

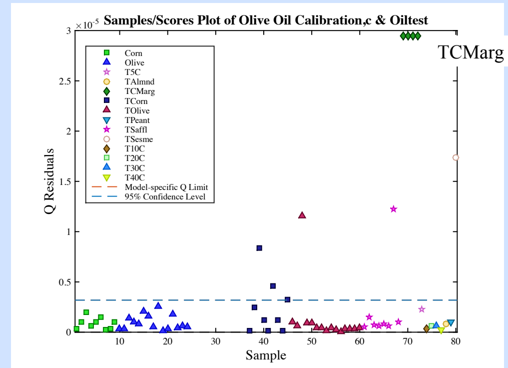
rows 10 to 24 correspond to olive





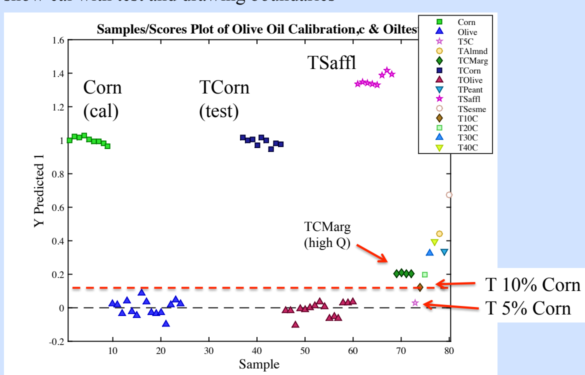
EIGENVECTOR RESEARCH INCORPORATED

using show cal with test and drawing boundaries



EIGENVECTOR RESEARCH INCORPORATED

CLS Results
using show cal with test and drawing boundaries

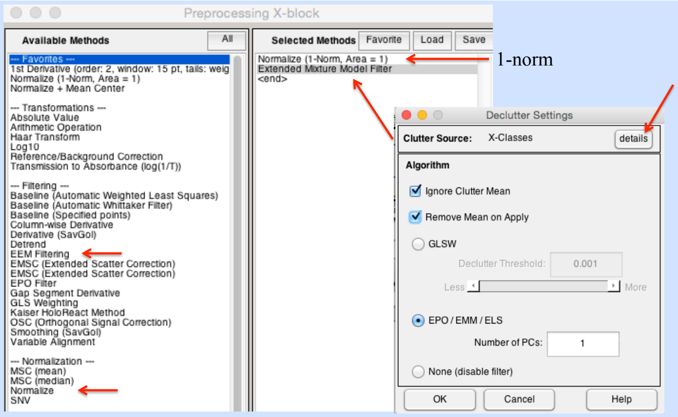


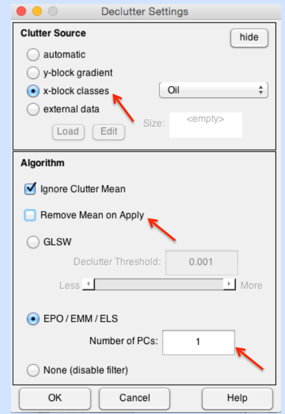
EIGENVECTOR RESEARCH INCORPORATED

CLS→ELS

- Learned that a large fraction (~80%) of the intra-class variance could be modeled with one principal component (1 PC)
  - one-norm, plus class-centering
- Use what you know
  - we know the classes for the calibration data
  - allows a model of intra-class clutter

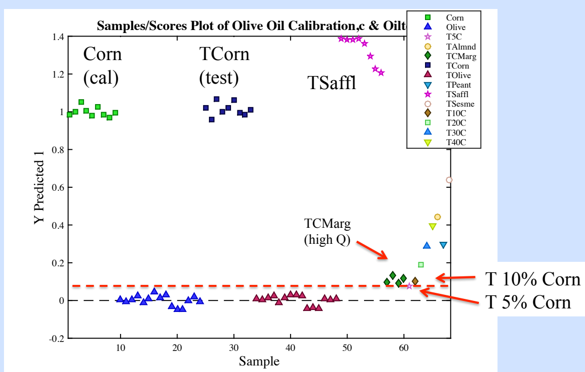
EIGENVECTOR RESEARCH INCORPORATED





1. click "x-block classes"
2. uncheck "Remove Mean on Apply"
3. Number of PCs = 1
4. ok

using show cal with test and drawing boundaries



Q Residuals for CLS & ELS

- ELS Q residuals are similar to those for PCA
  - The equation for a single measurement is

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

$$\mathbf{e} = \mathbf{x} - \begin{bmatrix} \mathbf{S} & \mathbf{P} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{c}}; & \hat{\mathbf{t}} \end{bmatrix}$$ Q contributions is a row of **E**

$$q = \mathbf{e}^T \mathbf{e}$$ Q residual is a sum-of-squared residuals

  - Limits for Q can be obtained using same tools used for PCA: Jackson, J.E. and Mudholkar, G.S., "Control Procedures for Residuals Associated with Principal Component Analysis," *Technometrics*, **21**(3), 341–349 (1979).

Limits for Scores

- Control limits can be placed on individual contributions and scores, **c** and **t** just like in PCA
  - Although they might not be normally distributed
- Limits might be set using statistical assumptions or engineering knowledge
  - e.g., control limits

Hotelling T^2 for CLS & ELS

- Hotelling T^2 is similar to that for PCA
  - The equation for a single measurement is

$$T^2 = \begin{bmatrix} \hat{\mathbf{c}}; & \hat{\mathbf{t}} \end{bmatrix}^T \left\{ \tfrac{1}{M-1} \begin{bmatrix} \mathbf{C} & \mathbf{T} \end{bmatrix}^T \begin{bmatrix} \mathbf{C} & \mathbf{T} \end{bmatrix} \right\}^{-1} \begin{bmatrix} \hat{\mathbf{c}}; & \hat{\mathbf{t}} \end{bmatrix}$$

C and **T** are for the calibration set

  - Limits for T^2 can be obtained using same tools used for PCA: Jackson, J.E., "A User's Guide to Principal Components", John Wiley & Sons, New York, NY (1991).

Comparison of CLS and ELS

- Because the basis for the interferences are augmented to the spectra, the math for CLS and ELS are identical.
- Therefore, w/o loss of generality ELS and CLS can be discussed under the general heading of "CLS."
- However, if we really need to split the pieces apart we can
  - … and the ELS approach also can be treated as a "weighted" CLS model as shown below

Weighted Least Squares, WLS

$$\mathbf{x} = \mathbf{c}\mathbf{S}^T + \mathbf{e}$$

Weighted least squares (WLS) model.

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

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

The residuals, **e** are assumed to be mean zero and have different variances for each entry. The residuals are assumed to be statistically independent.

$$\mathbf{e} \sim N\left(0, \boldsymbol{\sigma}^2\right)$$

$$\mathbf{e}\mathbf{W}^{-1/2} \sim N\left(0, \boldsymbol{\sigma}^2\mathbf{1}\right)$$

$$\mathbf{W} = diag\left(\boldsymbol{\sigma}^2\right)$$

Generalized Least Squares, GLS

$$\mathbf{x} = \mathbf{c}\mathbf{S}^T + \mathbf{e}$$

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

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

Generalized least squares (GLS) model.

The residuals, **e** are assumed to be mean zero and have different variances for each entry. The residuals are not assumed to be statistically independent.

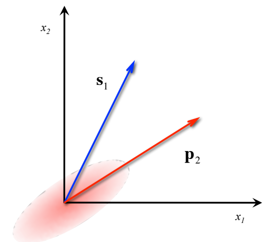
$$V(\mathbf{e}) = \mathbf{W}_c$$

$$\mathbf{e}\mathbf{W}_c^{-1/2} \sim N\left(0, \sigma^2\mathbf{1}\right)$$

68 Generalized Least Squares

EIGENVECTOR RESEARCH INCORPORATED

GLS in "Two-Space"



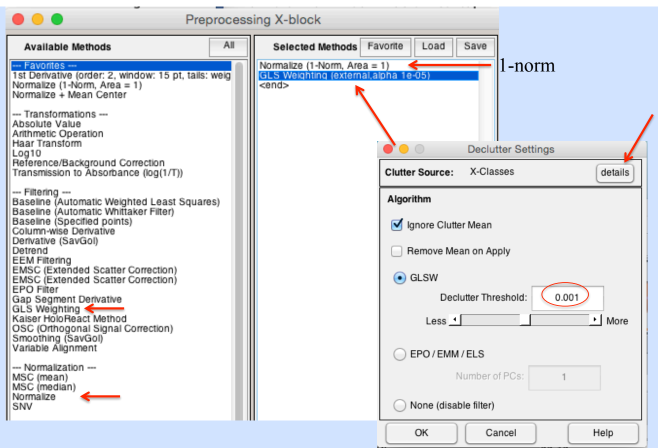
GLS attempts to model the clutter in a weighting matrix.

In the example shown here, the model might include both $\mathbf{s}_1$ and $\mathbf{p}_2$ as "spectra," as in ELS, while the fuzzy ball corresponds to the covariance of the clutter $\mathbf{W}_c$.

Choose the model structure appropriate for your data (learned from exploratory analysis).

69

EIGENVECTOR RESEARCH INCORPORATED

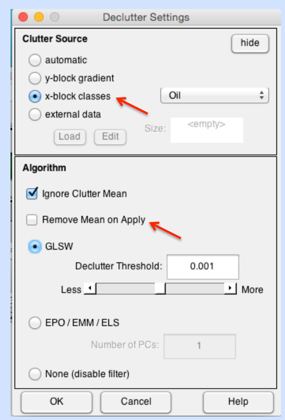


1-norm

GLS on Olive Oil

70 Generalized Least Squares-Olive Oil

EIGENVECTOR RESEARCH INCORPORATED

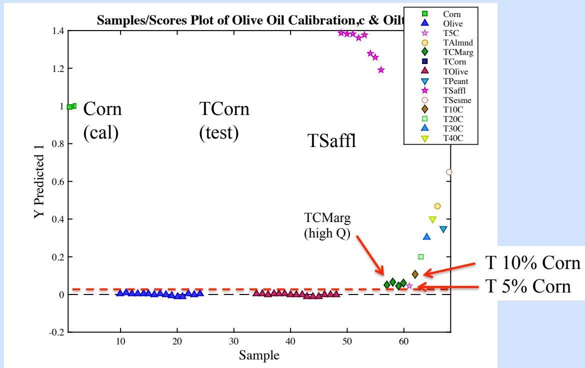


1. click "x-block classes"
2. uncheck "Remove Mean on Apply"
3. ok

71 Generalized Least Squares-Olive Oil

EIGENVECTOR RESEARCH INCORPORATED

using show cal with test and drawing boundaries



Plot Q and Y Predicted 2. Can change method options to use non-negative least squares.

EIGENVECTOR RESEARCH INCORPORATED

Classical Least Squares Comparison

$$\mathbf{x} = \mathbf{c}\mathbf{S}^T + \mathbf{e}$$

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

$$\mathbf{x} - \bar{\mathbf{x}} = \mathbf{c}\mathbf{S}^T + \mathbf{e}$$

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

$\mathbf{W} = \sigma^2\mathbf{I}$	CLS
$\mathbf{W} = \mathrm{diag}\left(\boldsymbol{\sigma}^2\right)$	WLS
$\mathbf{W} = \mathbf{W}_c$	GLS
$\mathbf{W}^{-1} = \left(\mathbf{I} - \mathbf{P}\left(\mathbf{P}^T\mathbf{P}\right)^{-1}\mathbf{P}^T\right)$	ELS

Mean-centering can be used to keep **e** mean zero.

EIGENVECTOR RESEARCH INCORPORATED

Orthogonalization and Weighting Flitering

- Comparison of CLS and Weighted CLS models

$$\mathbf{X} = \mathbf{C}\mathbf{S}^T \qquad \tilde{\mathbf{x}} = \mathbf{W}^{-1/2}\mathbf{x}$$

$$\tilde{\mathbf{S}} = \mathbf{W}^{-1/2}\mathbf{S}$$

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

Weighting by an inverse square root reduces the W-CLS model to CLS with weighted measurements and spectra i.e., the weighting can be viewed as a preprocessing step $\tilde{\mathbf{X}} = \mathbf{X}\mathbf{W}^{-1/2}$ that can be used w/ PCA and ILS models (PLS, PCR).

This leads to External Parameter Orthogonalization and GLS Weighting methods

EIGENVECTOR RESEARCH INCORPORATED

CLS Model Uses

- CLS is used when noise in each of the *N* measurements is similar.
- WLS is used when noise is different for each of the *N* measurements.
- GLS and ELS is used when the noise is correlated e.g., due to interferences.
  - Clutter = interferences + noise
  - GLS is a true weighting while ELS orthogonalizes completely to clutter directions

EIGENVECTOR RESEARCH INCORPORATED

GLS for Target Detection

- Target analyte has a stationary response
  - target response **s** is available $\hat{c} = \mathbf{x}\mathbf{W}^{-1}\mathbf{s}\left(\mathbf{s}^T\mathbf{W}^{-1}\mathbf{s}\right)^{-1}$
  - reference values **y** are not available (!)
- Backgrounds are highly variable
  - severe and highly variable interference signal
    - changes spatially (images) and/or temporarily (time-series)
    - difficult to account for
- The clutter **W** can be updated

EIGENVECTOR RESEARCH INCORPORATED

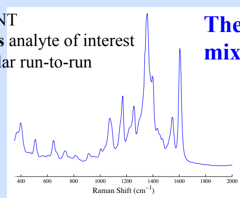
SERS Detection of DNT Vapor

- Detect trace (ppb/ppt) quantities of explosives vapor
  - "process" or "time-series" example
- Challenges
  - background / interferences change every run
    - difficult or impossible to span full background variability
  - multiple interferences
    - run-to-run and ubiquitous to every run

NB Gallagher[a], BD Piorek[b] ,SJ Lee[b], CD Meinhart[b], M Moskovits[b], BM Wise[a], "Multivariate Curve Resolution Applied to SERS Measurements of 2,4-DNT," APACT13 – 23-26 April, 2013
[a]Eigenvector Research, Inc.,
[b]SpectraFluidics, Inc.,

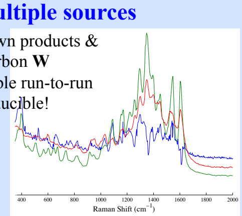
EIGENVECTOR RESEARCH INCORPORATED

2,4-DNT
target **s** analyte of interest
- similar run-to-run

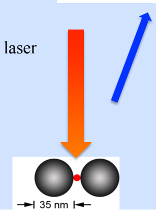


The measured signal is a mixture from multiple sources

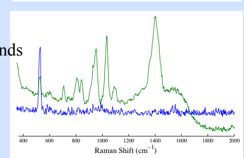
DNT breakdown products & amorphous carbon **W**
- highly variable run-to-run & not reproducible!



laser



Si background & citrate stabilizer **P**
- ubiquitous backgrounds



Ag Nanoparticle Aggregates
- Assemble around analyte molecules
- Provide SERS enhancement
- True concentration unknown

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

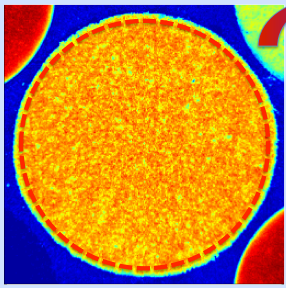
EIGENVECTOR RESEARCH INCORPORATED

Powdered Raw Materials

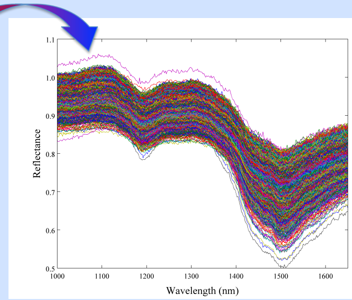
- Challenges – wheat gluten background
  - scattering and particle size distributions changes sample-to-sample
  - same material from a wide variety of sources
  - different powdered materials
- Calibration for typical ILS models difficult
  - reference values unavailable
  - unlikely to acquire a calibration data set that spans all the sample variation expected to be seen
    - and if you do, the net analyte signal suffers
  - unlikely to use one ILS model from a single material for multiple raw materials (other powders)

EIGENVECTOR RESEARCH INCORPORATED

Signal from the unadulterated wheat gluten is highly variable and is much stronger than the adulterant.

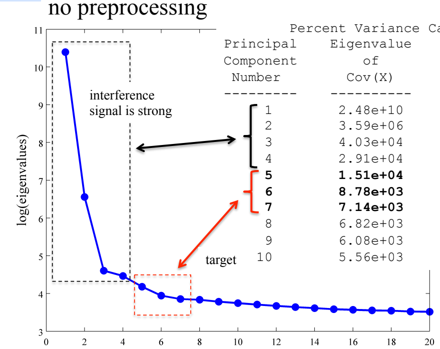


unadulterated wheat gluten



median background can be used as **p**

Eigenvalues for 200 ppm melamine in wheat gluten no preprocessing

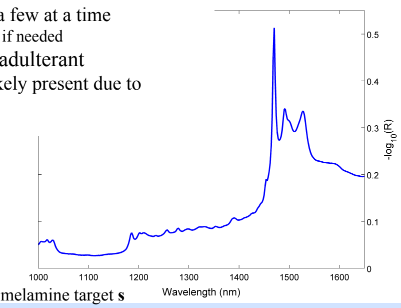


Percent Variance Captured by PCA Model

Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	2.48e+10	99.983438	99.983438
2	3.59e+06	0.014493	99.997931
3	4.03e+04	0.000162	99.998093
4	2.91e+04	0.000117	99.998210
5	**1.51e+04**	**0.000061**	**99.998271**
6	**8.78e+03**	**0.000035**	**99.998307**
7	**7.14e+03**	**0.000029**	**99.998336**
8	6.82e+03	0.000027	99.998363
9	6.08e+03	0.000025	99.998388
10	5.56e+03	0.000022	99.998410

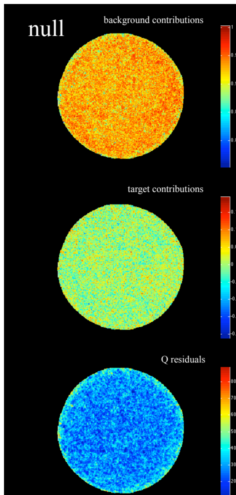
Adulterants = Targets

- Adulterant spectrum ~constant
  - variants can be considered a different adulterant
  - tend to appear with only a few at a time
    - multiples can be handled if needed
- Interest is in low levels of adulterant
  - however, higher levels likely present due to economics



melamine target **s**

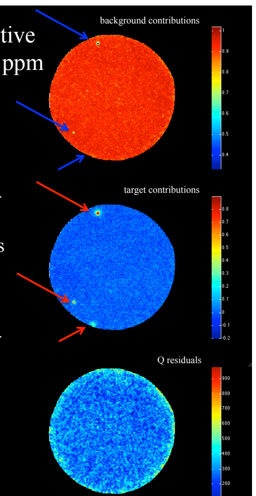
null



positive 200 ppm

example contributions for two images

1) response is not linear with ppm (mass) because only particles on top are detected
2) distributions are not normal
3) response is not linear in a pixel due to ~signal closure



GLS = Adaptive Matched Filter

- The GLS estimator is used for target detection in remote /standoff sensing and is referred to as "the matched filter."
  - ground truth is rarely know well
  - interferences vary most every measurement
- More recently as "the adaptive matched filter" because the clutter is updated.
- Understanding the source of clutter and how to account for mathematically is important

EIGENVECTOR RESEARCH INCORPORATED

GLS and Clutter

- The clutter covariance is estimated from target-free measurements or measurements where the target contributions do not change.*

For clutter that has a constant (stationary) mean, the clutter covariance is estimated from

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

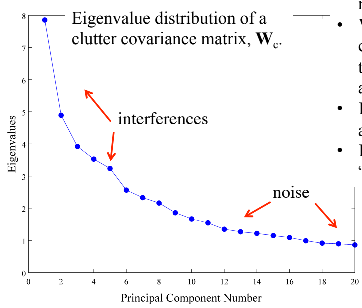
- Sensor noise and signal due to interferences
  - (DOD, NATO) Unwanted signals, echoes, or images on the face of the display tube, which interfere with observation of desired signals.
  - It's measured signal unrelated to the target
    - it can be correlated or not

*Typical use. Often target is also present – makes detection thresholds difficult to quantify

EIGENVECTOR RESEARCH INCORPORATED

Clutter Covariance Model

- How many PCs?



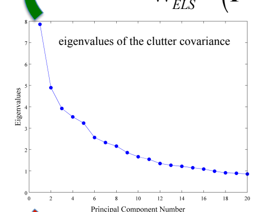
- In PCA, the big eigenvalues are kept to model the signal of interest in the data.
- When the data set corresponds to clutter, the big eigenvalues correspond to interference signal that has to be accounted for.
- In GLS, these are the directions that are down-weighted the most.
- In ELS, these are the directions that are "thrown away".

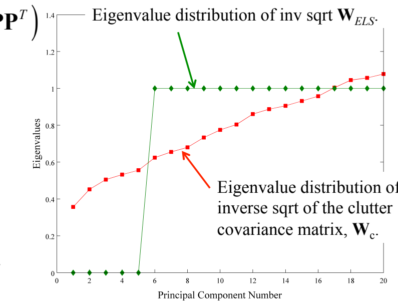
EIGENVECTOR RESEARCH INCORPORATED

Compare ELS and GLS

ELS: remove $\mathbf{p}_{1\text{-}5}$ of $\mathbf{W}_c$.

$$\mathbf{W}_{ELS}^{-1/2} = \left(\mathbf{I} - \mathbf{P}\mathbf{P}^T\right)$$





GLS: take inverse of $\mathbf{W}_c$.

$$\mathbf{W}_c = \mathbf{P}\mathbf{S}_c\mathbf{P}^T$$

$$\mathbf{W}_c^{-1/2} = \mathbf{P}\mathbf{S}_c^{-1/2}\mathbf{P}^T$$

EIGENVECTOR RESEARCH INCORPORATED

Compare ELS and GLS

- ELS is a hard cutoff while GLS is a "soft" cutoff.
  - GLS doesn't throw away intermediate clutter eigenvalues
- ELS assumes that the clutter eigenvalues of the kept subspace are all the same.
  - ELS assumes statistically independent residuals of similar magnitude
    - a diagonal matrix with all entries the same value has a flat eigenvalue distribution
    - if truly statistically independent, the eigenvalues are std($\mathbf{X}$)^2
- If the eigenvalue distribution of the clutter covariance is flat, GLS will not de-weight any directions
  - estimated covariance matrices rarely have a flat eigenvalue distribution

EIGENVECTOR RESEARCH INCORPORATED

Constraints

- Constraints and penalties add control over the estimator
- The advantage for CLS is that
  - the objective function is based on estimating contributions or concentrations **c**,
  - not on estimating the regression vector ***b*** as with ILS methods

EIGENVECTOR RESEARCH INCORPORATED

Non-negativity

- Physics and chemistry often dictate the contributions must be non-negative. This can be added as a constraint to the least-squares solution
- force $\hat{\mathbf{c}}$ to be ≥ 0
  - or $\geq$ a small tolerance (e.g., slightly <0 due to noise)
- not all contributions need be non-negative

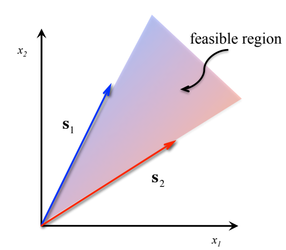
$$\hat{\mathbf{c}} = \mathbf{x}\mathbf{W}^{-1}\mathbf{S}\left(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S}\right)^{-1}$$

e.g., non-negativity for ELS contributions on **P** might be relaxed

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

EIGENVECTOR RESEARCH INCORPORATED

Non-negativity in "Two-Space"

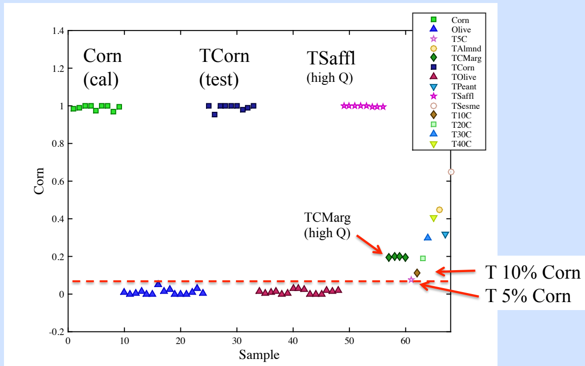


Non-negativity forces all measurements to lie between $\mathbf{s}_1$ and $\mathbf{s}_2$.

Can also set a tolerance allowing the signal to be slightly outside the feasible region.

EIGENVECTOR RESEARCH INCORPORATED

using show cal with test and drawing boundaries



change method options to use non-negative least squares
manually set nconst = [1 1 0]

101 Extended Least Squares-Olive Oil

EIGENVECTOR RESEARCH INCORPORATED

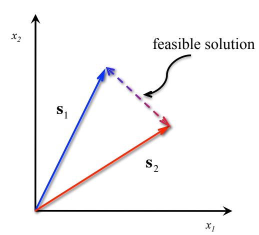
Closure

- Physics and chemistry often dictate the contributions must sum to one (i.e., obey closure). This can be added as a constraint to the least-squares solution
- forces $\sum_{k=1}^{K} c_i = 1$
- Is used in combination with non-negativity
- not all contributions need obey closure
  - e.g., closure for ELS contributions on **P** might be relaxed

102 Penalties and other Constraints

EIGENVECTOR RESEARCH INCORPORATED

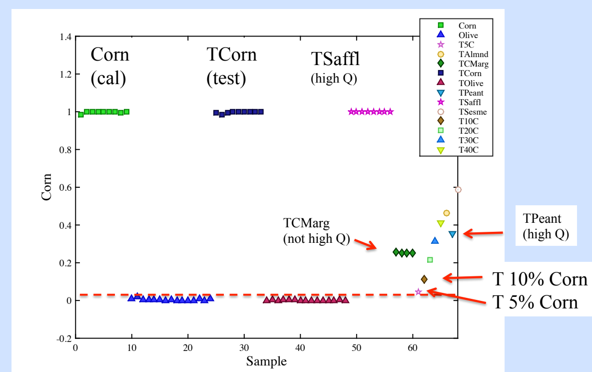
Closure in "Two-Space"



Closure implies non-negativity and forces all measurements to lie on a line between $\mathbf{s}_1$ and $\mathbf{s}_2$.

103

EIGENVECTOR RESEARCH INCORPORATED



change method options to use non-negative least squares
manual application of closure

105 Generalized Least Squares-Olive Oil

EIGENVECTOR RESEARCH INCORPORATED

CLS w/ a Prior + Smooth

$$\mathbf{x} = \mathbf{Sc} + \mathbf{e}$$

$$O(\mathbf{c}) = (\mathbf{x} - \mathbf{Sc})^T \mathbf{W}^{-1}(\mathbf{x} - \mathbf{Sc}) + \alpha_1(\mathbf{c} - \mathbf{c}_0)^T \mathbf{A}^{-1}(\mathbf{c} - \mathbf{c}_0) + \alpha_2 \mathbf{c}^T \mathbf{D}^T \mathbf{Dc}$$

weighted residuals
penalty function to introduce a prior
penalty function to introduce smoothing

A is a covariance for **c**, and **D** might be a 2nd derivative operator to introduce a penalty on roughness (it introduces smoothness). As α_2 gets large **c** gets more smooth.

$$\hat{\mathbf{c}} = \left(\mathbf{S}^T\mathbf{S} + \alpha_1\mathbf{A}^{-1} + \alpha_2\mathbf{D}^T\mathbf{D}\right)^{-1}\left(\mathbf{S}^T\mathbf{x} + \alpha_1\mathbf{A}^{-1}\mathbf{c}_0\right)$$

note that **D** need not be continuously banded

CLS w/ Basis Functions

- Assume **c** = **Bb** where **B** is a set of basis functions known *a priori* (e.g., splines, spectra, PCs or other) and **b** is the set of coefficients to identify.
  - It is typical that the number of basis functions in **B** is smaller than the number of spectra in **S**.
  - Approach can be used to employ smoothing.
  - Development formalization can be used to derive PCR.

$$\mathbf{c} = \mathbf{Bb} \quad ; \quad \hat{\mathbf{c}} = \mathbf{B}\hat{\mathbf{b}}$$

$$\mathbf{x} = \mathbf{Sc} + \mathbf{e}$$

$$O(\mathbf{c}) = (\mathbf{x} - \mathbf{Sc})^T \mathbf{W}^{-1}(\mathbf{x} - \mathbf{Sc}) + \alpha_1 \mathbf{c}^T \mathbf{A}^{-1}\mathbf{c}$$

$$O(\mathbf{b}) = (\mathbf{x} - \mathbf{SBb})^T \mathbf{W}^{-1}(\mathbf{x} - \mathbf{SBb}) + \alpha_1 \mathbf{b}^T \mathbf{B}^T \mathbf{A}^{-1}\mathbf{Bb}$$

$$\hat{\mathbf{b}} = \left(\mathbf{B}^T\mathbf{S}^T\mathbf{W}^{-1}\mathbf{SB} + \alpha_1\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{S}^T\mathbf{x}$$

$$\hat{\mathbf{c}} = \mathbf{B}\left(\mathbf{B}^T\mathbf{S}^T\mathbf{W}^{-1}\mathbf{SB} + \alpha_1\mathbf{B}^T\mathbf{A}^{-1}\mathbf{B}\right)^{-1}\mathbf{B}^T\mathbf{S}^T\mathbf{x}$$

Net Analyte Signal

- Net analyte signal (NAS)
  - The portion of signal unique to each analyte
    - it is the part of $\mathbf{s}_k$ orthogonal to interferences
- For a generality, NAS is defined for an ELS model as

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

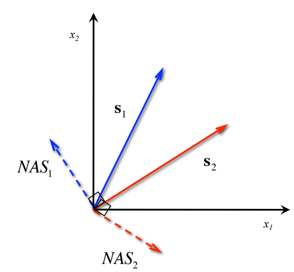
ELS model:

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

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

A. Lorber, B. Kowalski, "The Effect of interferences and Calibration Design on Accuracy: Implications for Sensor and Sample Selection," *J. Chemom.*, **2**, *67-79* (1988)

NAS in "Two-Space"



NAS is the portion of the signal unique to each analyte. NAS_1 is the portion of $\mathbf{s}_1$ parallel to $\mathbf{s}_1$ and orthogonal to $\mathbf{s}_2$.

Outline / Summary

- Linear Mixture Model
  - Classical Least Squares (CLS), Weighted Least Squares (WLS), Extended Least Squares (ELS), Generalized Least Squares (GLS)
    - all can be contained in a WLS framework
    - orthogonalization and weighting filters
    - models are interpretable
- Target Detection
    - used when target spectrum available but no reference values
- Concept of Clutter
    - accounting for interferences in the data (use what you know)
    - models are easy to update
- Constraints
    - constraints on what is estimated, **c** (not on a regression vector)
    - added control over modeling (use what you know)
- Net Analyte Signal

113

EIGENVECTOR RESEARCH INCORPORATED

Keep in mind that…

"The detection, classification and/or quantification system being considered is the 1) sensor that provides the measurements, 2) the scenario in which it is to be deployed *and* 3) the algorithm used to extract the desired information.

These must be developed concurrently for the greatest chance at success because what is learned during data analysis and algorithm development often feeds back directly to instrument design in an effort to maximize
signal-to-clutter not just *signal-to-noise*."

EIGENVECTOR RESEARCH INCORPORATED

CLS Regression Methods

(Building Interpretable Predictive Models)

Appendix
Advanced Examples and
Additional Concepts

116

EIGENVECTOR RESEARCH INCORPORATED

How to use MCR to get P?

- When less is known a about the data…
- Given known **X** and **C**, how can we estimate **S**, **P** and **T** from a calibration set?
  - **S** and **P** can be used to make estimates for **C** (and **T**) from a test set.

target factors　　interferences

$$\mathbf{X} = \begin{bmatrix} \mathbf{C} & \mathbf{T} \end{bmatrix} \begin{bmatrix} \mathbf{S} & \mathbf{P} \end{bmatrix}^T + \mathbf{E}$$

117 Extended Least Squares

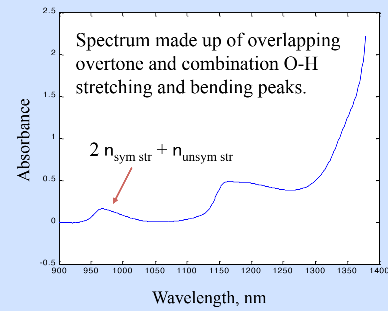
EIGENVECTOR RESEARCH INCORPORATED

Caustic Data Example

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

EIGENVECTOR RESEARCH INCORPORATED

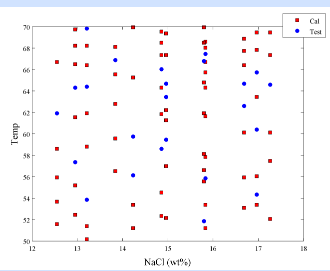
Typical SW-NIR Spectrum of Caustic Brine Solution

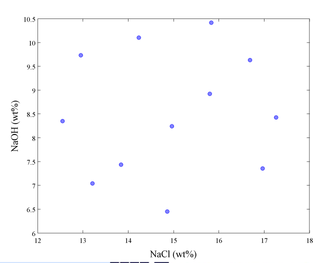


EIGENVECTOR RESEARCH INCORPORATED

DOE for Calibration

- NaOH, NaCl and T varied in a designed experiment.
- Split the data into calibration and test (are they independent sets?)
  - 71 calibration spectra (red)
  - 24 test spectra (blue)





What is the anticipated rank of the calibration set?

EIGENVECTOR RESEARCH INCORPORATED

Preprocessing

- Second Derivative (15,5,2)
- Mean-center
  - often with spectra no centering is used
    - no centering is a force fit through zero
  - mean-centering is used to introduce an offset
    - often used with ILS models
- Use same preprocessing for ELS and PLS so a fair comparison can be made

EIGENVECTOR RESEARCH INCORPORATED

Problem Summary

known

X, measured spectra calibration set
C, concentration of NaCl and NaOH

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

unknown

S, spectra of NaCl and NaOH
T, interference "contributions" or scores
P, basis for interferences but can be estimated - use data for constant NaCl and NaOH and varying temp
E, residuals

EIGENVECTOR RESEARCH INCORPORATED

MCR to Estimate the ELS Model

known

C, concentration of NaCl and NaOH, constrained "softly"

P, estimated from samples at constant **c** and changing T, constrained "softly"

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

unknown

S, spectra of NaCl and NaOH unconstrained
T, interference contributions unconstrained

recall derivatives and mean-centering means that contributions and spectra can be negative

EIGENVECTOR RESEARCH INCORPORATED

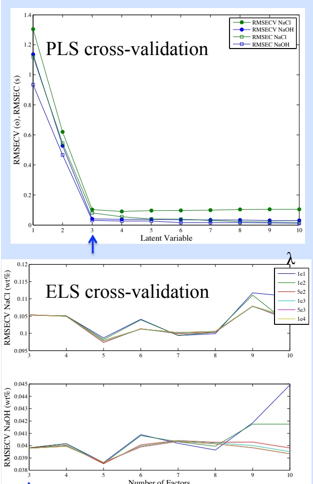
Soft Constraints for C and S

- Augment to **X** "adds weight" to the "knowns"
  - large weight tends to "hard" constrain
  - cross-validate over the weight λ
  - cross-validate over the number of factors in **P**

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

$$\begin{bmatrix} \mathbf{X} \\ \mathbf{P}\lambda \end{bmatrix} = [\tilde{\mathbf{C}}\ \ \tilde{\mathbf{T}}][\mathbf{S}\ \ \mathbf{P}]^T + \mathbf{E}$$

EIGENVECTOR RESEARCH INCORPORATED



Results

Examine two different models:

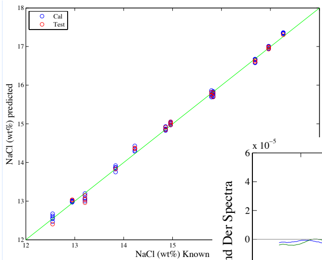
3 factors for PLS and ELS (λ=500)

(wt%)	NaCl PLS	NaCl ELS	NaOH PLS	NaOH ELS
RMSEC	0.080	0.083	0.031	0.032
RMSECV	0.103	0.097	0.041	0.038
RMSEP	0.099	0.099	0.032	0.034

5 factors for PLS and ELS (λ=500)

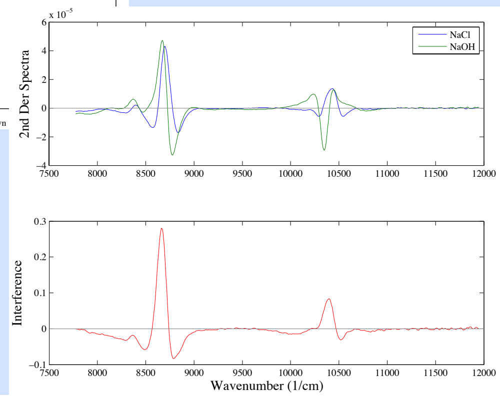
(wt%)	NaCl PLS	NaCl ELS	NaOH PLS	NaOH ELS
RMSEC	0.039	0.075	0.026	0.029
RMSECV	0.096	0.100	0.037	0.040
RMSEP	0.068	0.095	0.029	0.034

EIGENVECTOR RESEARCH INCORPORATED



Results for ELS

Using 3 Factors

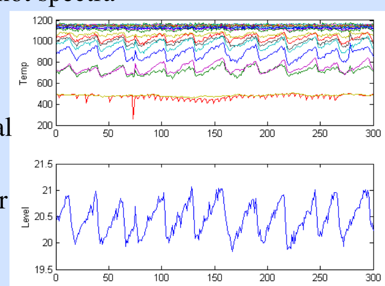


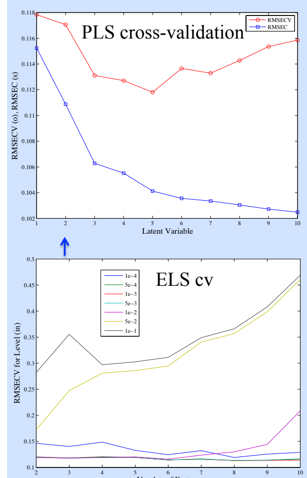
Summary: Caustic

- PLS performed ~slightly better than ELS
  - statistically significant?
  - algorithms for ELS not fully optimized and not clear how to do cross-validation
    - e.g., for soft constraints
    - SLOW! mostly due to ALS algorithm
  - data were from a good DOE so results were expected to be similar
    - not yet shown how to account for confounding in the measurements, but …
  - we have shown that a CLS model can be used even when the spectrum of the interferent was unknown

SFCM Data Example

- Estimate level in a slurry fed ceramic melter
  - measurements are not spectra
  - measured 20 temperatures (thermocouples) in two vertical thermal wells
  - thermocouples near the surface vary with level

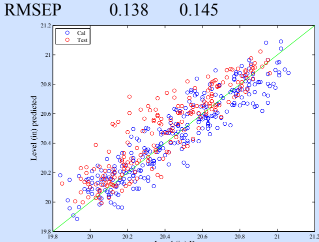


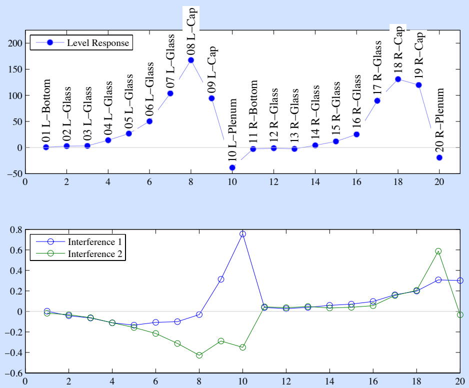


Results

Demonstration of CLS for engineering variables.
3 factors for PLS and ELS (λ=0.001)

	Level (in)	
	PLS	ELS
RMSEC	0.106	0.114
RMSECV	0.113	0.118
RMSEP	0.138	0.145





EIGENVECTOR RESEARCH INCORPORATED

Derivation of EPO Preprocessing

Start with the ELS model and show that the regression vector is the same as the EPO-based regression vector.

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

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

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_s & \mathbf{B}_p \end{bmatrix} = \begin{bmatrix} \mathbf{S} & \mathbf{P} \end{bmatrix} \left(\begin{bmatrix} \mathbf{S} & \mathbf{P} \end{bmatrix}^T \begin{bmatrix} \mathbf{S} & \mathbf{P} \end{bmatrix} \right)^{-1}$$

$$\begin{bmatrix} \mathbf{S} & \mathbf{P} \end{bmatrix}^T \begin{bmatrix} \mathbf{S} & \mathbf{P} \end{bmatrix} = \begin{bmatrix} \mathbf{S}^T\mathbf{S} & \mathbf{S}^T\mathbf{P} \\ \mathbf{P}^T\mathbf{S} & \mathbf{P}^T\mathbf{P} \end{bmatrix}$$

EIGENVECTOR RESEARCH INCORPORATED

Estimation Error for CLS

- The easy derivation is also likely the most important in general.
  - Assumes that the measurement error in **x** dominates error in **S** (or [**S P**] for ELS).
    - neglects the leverage term
    - describes error variance about the model origin
      - useful for estimating detection thresholds
    - often true when **S** estimated from high quality lab
  - Not always true because **S** and **P** are often estimated directly from measurements
    - library **S** not available or not exactly problem-relevant

EIGENVECTOR RESEARCH INCORPORATED

Estimation Error

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

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

$$\mathrm{V}_{CLS}(\mathbf{x}) = \mathbf{W} = \sigma^2\mathbf{I} \qquad \mathrm{V}_{CLS}(\hat{\mathbf{c}}) = \sigma^2\left(\mathbf{S}^T\mathbf{S}\right)^{-1}$$

$$\mathrm{V}_{WLS}(\mathbf{x}) = \mathbf{W} = \mathrm{diag}\left(\boldsymbol{\sigma}^2\right) \qquad \mathrm{V}_{WLS}(\hat{\mathbf{c}}) = \left(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S}\right)^{-1}$$

$$\mathrm{V}_{GLS}(\mathbf{x}) = \mathbf{W} = \mathbf{W}_c \qquad \mathrm{V}_{GLS}(\hat{\mathbf{c}}) = \left(\mathbf{S}^T\mathbf{W}_c^{-1}\mathbf{S}\right)^{-1}$$

$$\mathrm{V}_{ELS}(\mathbf{x}) = \sigma^2\mathbf{I} \qquad \mathrm{V}_{ELS}\left(\begin{bmatrix} \hat{\mathbf{c}} & \hat{\mathbf{t}} \end{bmatrix}\right) = \sigma^2\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}$$

where $\mathbf{Z} = \begin{bmatrix} \mathbf{S} & \mathbf{P} \end{bmatrix}$

EIGENVECTOR RESEARCH INCORPORATED

NAS Compared to Estimation Error

$$\mathbf{b}_k = \left[\mathbf{I} - \mathbf{P}\left(\mathbf{P}^T\mathbf{P}\right)^{-1}\mathbf{P}^T\right]\mathbf{s}_k\left(\mathbf{s}_k^T\left[\mathbf{I} - \mathbf{P}\left(\mathbf{P}^T\mathbf{P}\right)^{-1}\mathbf{P}^T\right]\mathbf{s}_k\right)^{-1} \qquad \text{NAS}$$

$$\mathbf{b}_k^T\mathbf{b}_k = \left(\mathbf{s}_k^T\left[\mathbf{I} - \mathbf{P}\left(\mathbf{P}^T\mathbf{P}\right)^{-1}\mathbf{P}^T\right]\mathbf{s}_k\right)^{-1}\mathbf{s}_k^T\left[\mathbf{I} - \mathbf{P}\left(\mathbf{P}^T\mathbf{P}\right)^{-1}\mathbf{P}^T\right]\mathbf{s}_k\left(\mathbf{s}_k^T\left[\mathbf{I} - \mathbf{P}\left(\mathbf{P}^T\mathbf{P}\right)^{-1}\mathbf{P}^T\right]\mathbf{s}_k\right)^{-1}$$

$$\mathbf{b}_k^T\mathbf{b}_k = \left(\mathbf{s}_k^T\left[\mathbf{I} - \mathbf{P}\left(\mathbf{P}^T\mathbf{P}\right)^{-1}\mathbf{P}^T\right]\mathbf{s}_k\right)^{-1}$$

is the the scalar diagonal element for the k^{th} analyte of $\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}$
that shows that the longer the NAS, the smaller the error because
the estimation error is $\sigma^2\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}$