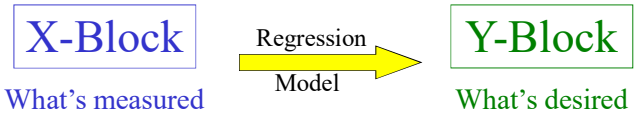
Chemometrics II: Regression and PLS

(Building Predictive Models)

EIGENVECTOR RESEARCH INCORPORATED

Regression



Regression analysis creates a mapping between two blocks of data.

In contrast, PCA was used to explore the correlation structure within a single data block.

Regression models are often used to obtain estimates (or predictions) for one block of data from the other.

EIGENVECTOR RESEARCH INCORPORATED

What Can Be Done with Regression?

- Analyte concentrations from spectra or other sensors
  - CH_4, H_2O, CO_2 in natural gas (NIR)
  - H_2, NH_3 in waste tanks (FTIR)
  - Sugar content of fruit (NIR)
- Prediction of property values
  - Octane of gasoline (NIR)
  - Ozone forming potential of automobile exhaust (FTIR)
- Sample classification (e.g., PLS-DA)
  - Detection of cervical cancer (ETF)
  - Detection of atherosclerotic (vulnerable) plaques (NIR)

EIGENVECTOR RESEARCH INCORPORATED

Outline

EIGENVECTOR RESEARCH INCORPORATED

Course Materials

- These slides
- PLS_Toolbox or Solo 8.1 or later
- Data sets
  - From DEMS folder (installed with software)
    - plsdata (SFCM)
  - From EVRIHW folder (additional data sets)
    - EigenU_nir_data, SBRdata_EU

Conventions & Notation

- *Rows* correspond to *samples*, *columns* correspond to *variables*
- Notation:
  - **X** = matrix of predictor variables
  - **Y** = matrix (or vector **y**) of predicted variables
  - M = number of samples (observations)
  - N_x = number of **X** variables, N_y = number of **Y** variables
  - **T** = **X**-block scores matrix, $\mathbf{t}_1$, $\mathbf{t}_2$, ..., $\mathbf{t}_K$ score vectors
  - **U** = **Y**-block scores matrix, $\mathbf{u}_1$, $\mathbf{u}_2$, ..., $\mathbf{u}_K$ score vectors
  - **P** = **X**-block loads matrix, $\mathbf{p}_1$, $\mathbf{p}_2$, ..., $\mathbf{p}_K$ loadings vectors
  - **Q** = **Y**-block loads matrix, $\mathbf{q}_1$, $\mathbf{q}_2$, ..., $\mathbf{q}_K$ loadings vectors
  - **W** = **X**-block weights matrix, $\mathbf{w}_1$, $\mathbf{w}_2$, ..., $\mathbf{w}_K$ loadings vectors
  - Θ = ridge parameter

Data Preprocessing

- Everything that was said about preprocessing for PCA goes double for regression
- Data should be linearized, if possible
- Data is often mean-centered
- Variance scaling used when variables are in different units or greatly different magnitudes
- Many preprocessing methods available!
  - Goal: reduce extraneous variance, emphasize relevant variance
- Outlier elimination is critical to regression models

Classical Least Squares

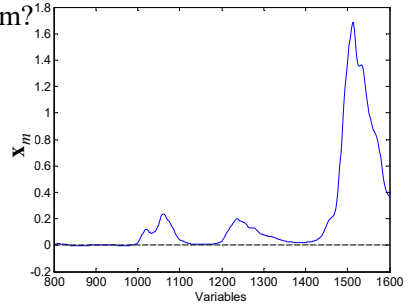
- CLS can be used to develop calibration models
  - often used in spectroscopy
- The CLS model assumes the data follows:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^{\mathrm{T}} + \mathbf{E}$$

where **X** (MxN_x) is the measured response, **S** (N_xxK) is a matrix of pure component responses, **C** (MxK) is a matrix of weights (*e.g.*, concentrations) and **E** (MxN_x) is noise or an error matrix.

The CLS Model

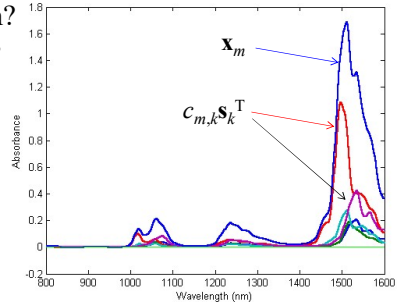
- Given known pure component spectra, how much of each does it take to make up the observed m^{th} spectrum?
- $\mathbf{x}_m = \mathbf{c}_m\mathbf{S}^T + \mathbf{e}_m$
- $m = 1,\ldots,M$
- $\mathbf{c}_m=[c_{m,1}, c_{m,2}, \ldots, c_{m,K}]$
- $k = 1,\ldots,K$



EIGENVECTOR RESEARCH INCORPORATED

The CLS Model

- Given known pure component spectra **S**, how much of each does it take to make up the observed spectrum?
- *i.e.,* what are the $c_{m,k}$?



EIGENVECTOR RESEARCH INCORPORATED

CLS (cont.)

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

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

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

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

- Problem: How to get **S**?
  - library, estimate from calibration measurements

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

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:

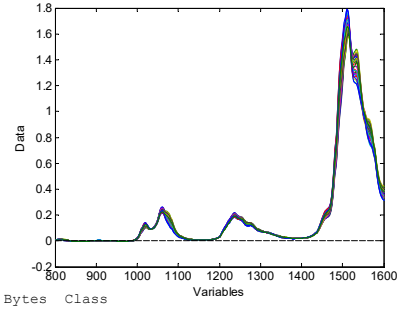
$$\mathbf{S}^T = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{X}$$

  - 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.

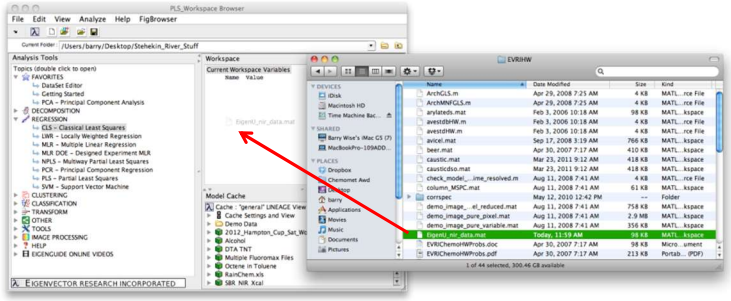
CLS Example

- 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

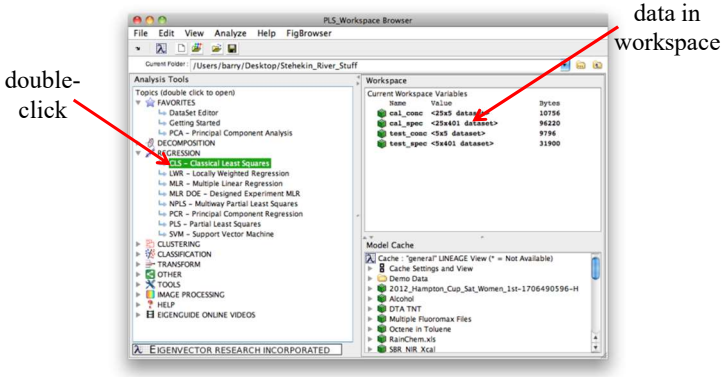



```
>> load EigenU_nir_data
>> whos
  Name            Size              Bytes  Class
  cal_conc        25x5              11002  dataset
  cal_spec        25x401            96466  dataset
  test_conc       5x5               10042  dataset
  test_spec       5x401             32146  dataset
```

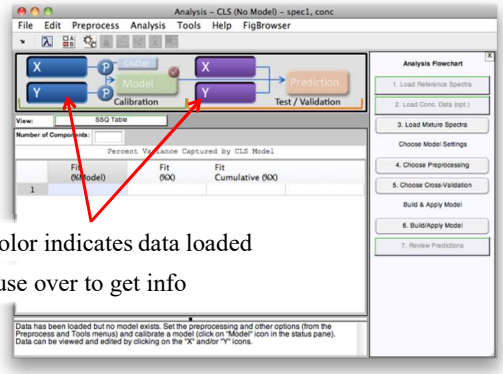

Load Data Into Browser



Start CLS Interface



Data Loaded

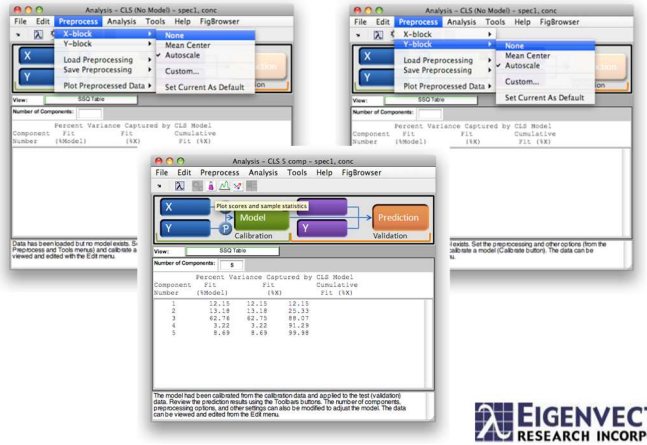


darker color indicates data loaded

mouse over to get info

EIGENVECTOR RESEARCH INCORPORATED

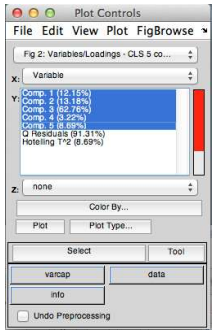
Set Preprocessing to "none," calculate model



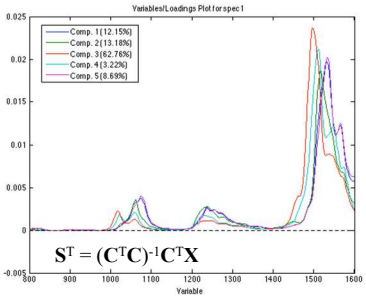
EIGENVECTOR RESEARCH INCORPORATED

Pure Component Spectra

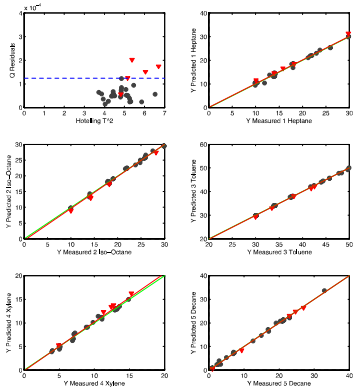
Click loadings "spectrum" icon, select all 5 components



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.

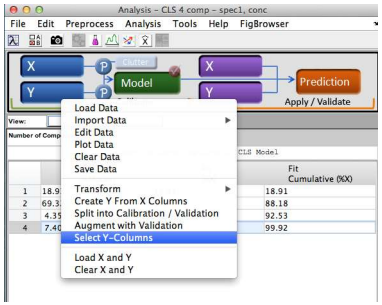
EIGENVECTOR RESEARCH INCORPORATED

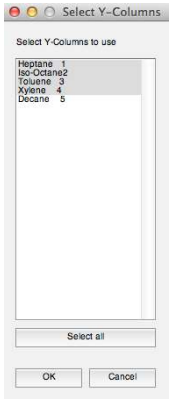
CLS Problem

- 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

EIGENVECTOR RESEARCH INCORPORATED

Select only the first four analytes and repeat

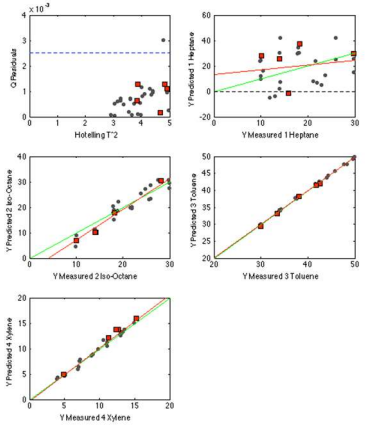




click 'cal Y: select Y-columns'

EIGENVECTOR RESEARCH INCORPORATED

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

EIGENVECTOR RESEARCH INCORPORATED

Inverse Least Squares

- Inverse least squares (ILS) models assume that the model is of the form:

$$\mathbf{Xb} = \mathbf{y} + \mathbf{e}$$

where $\mathbf{y}$ (Mx1) is a property to be predicted, $\mathbf{X}$ (MxN_x) is the measured response, $\mathbf{e}$ (Mx1) is an error vector, and $\mathbf{b}$ (N_xx1) is a vector of coefficients

- Unlike CLS, ILS methods associate the noise with the predicted property, not the measured response

EIGENVECTOR RESEARCH INCORPORATED

Advantage of ILS Methods

- ILS methods (including MLR, PCR, PLS, CR) don't require the concentration of all analytes, including interferents, be known …
- …however, interferents must vary in the calibration data set for the ILS regression model to be robust against them

Interferent: Any substance whose presence interferes with an analytical procedure and generates incorrect results (wiktionary)

Estimation of b: MLR

- It is possible to estimate **b** from

$$\mathbf{b} = \mathbf{X}^{+}\mathbf{y}$$

where $\mathbf{X}^{+}$ is the pseudo-inverse of $\mathbf{X}$

- There are many ways to obtain a pseudo-inverse; the most obvious is multiple linear regression (MLR), a.k.a., Ordinary Least Squares (OLS)
- In this case, $\mathbf{X}^{+}$ is estimated from

$$\mathbf{X}^{+} = \left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T}$$

Multiple Linear Regression

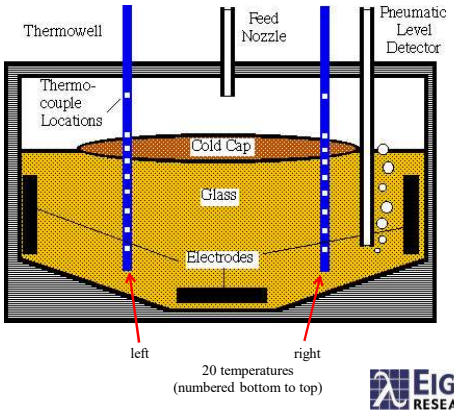
$$\mathbf{Xb} = \mathbf{y} + \mathbf{e}$$
$$\mathbf{Xb} = \mathbf{y}$$
$$\mathbf{X}^{T}\mathbf{Xb} = \mathbf{X}^{T}\mathbf{y}$$
$$\mathbf{b} = \left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T}\mathbf{y}$$
$$\mathbf{X}^{+} = \left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T}$$

- Note that $\mathbf{X}^{T}\mathbf{X}$ is N_xxN_x and square

Problem with MLR

- Inverse of $\mathbf{X}^{T}\mathbf{X}$ only exists if …
  - Rank(**X**) = N_x, however rank(**X**) ≤ min (M,N_x)
  - **X** has more samples than variables *i.e.,* if M>N_x, and
    - problem with spectra
  - Columns of **X** are not co-linear.
- Inverse may exist but be highly unstable if **X** is nearly rank deficient (a.k.a., ill-conditioned).
- In these cases, small perturbations in the data (possibly due to noise) can produce very different results.

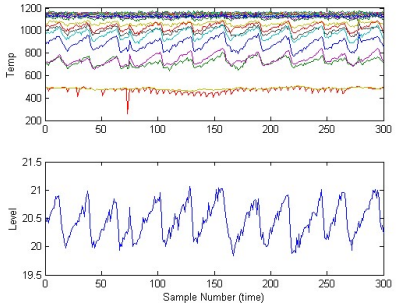
Slurry Fed Ceramic Melter: SFCM



EIGENVECTOR RESEARCH INCORPORATED

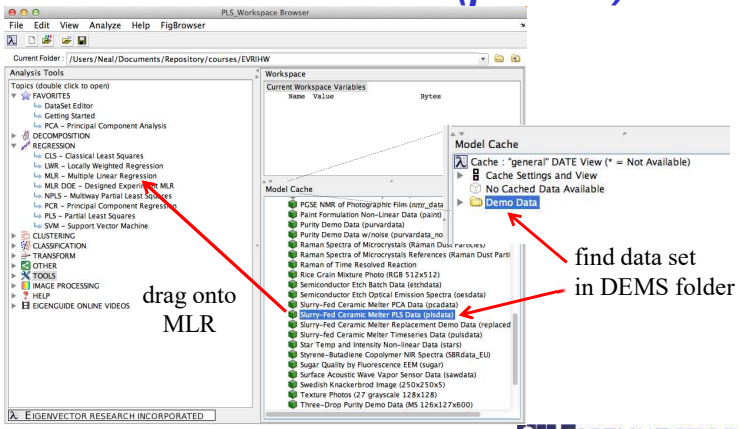
MLR Example

- Use MLR to obtain a relationship between temperature and level in a SFCM



EIGENVECTOR RESEARCH INCORPORATED

Load SFCM Data (plsdata)



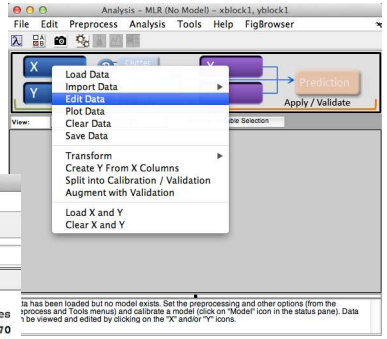
EIGENVECTOR RESEARCH INCORPORATED

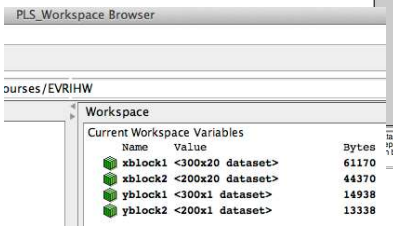
Edit Data

`xblock1` and `yblock1` will load as calibration

`xblock2` and `yblock2` will load as validation set

Edit calibration X-block data



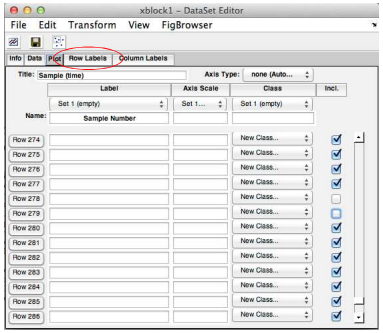


EIGENVECTOR RESEARCH INCORPORATED

Remove Outliers

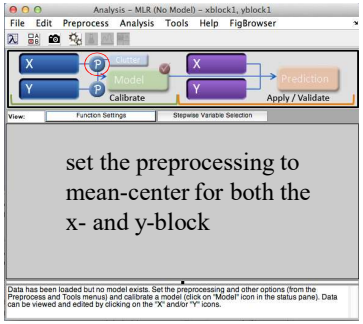
Select "Row Labels" tab in DataSet Editor

Exclude samples 73, 167, 278 and 279 from `xblock1`

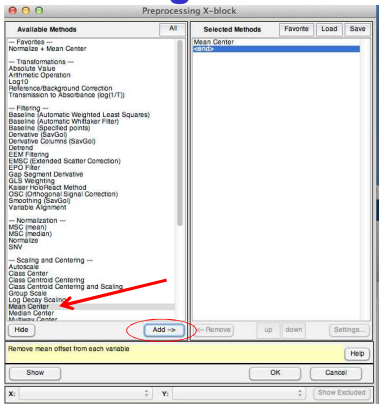


EIGENVECTOR RESEARCH INCORPORATED

Set Preprocessing

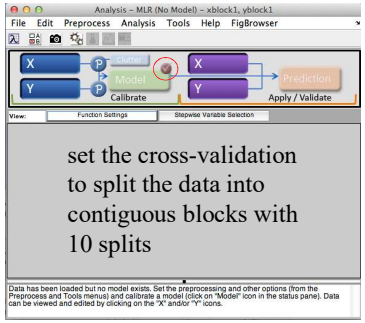


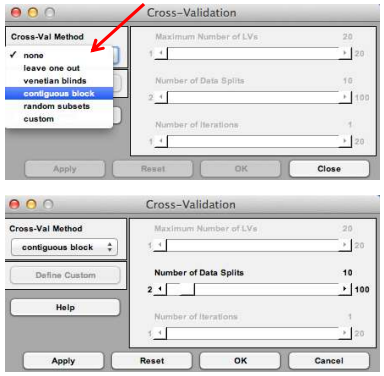
Mean-centering allows for an offset in the linear regression. Not centering is a force fit through zero.



EIGENVECTOR RESEARCH INCORPORATED

Set Cross-Validation



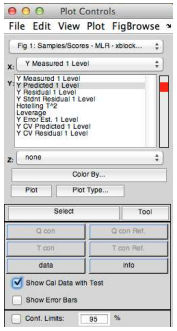


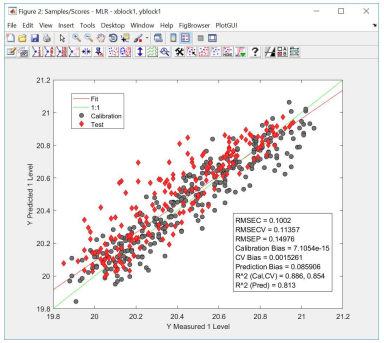
EIGENVECTOR RESEARCH INCORPORATED

MLR Fit to Calibration and Prediction on Validation Data

Click calculate model 

Click scores "flask" 





EIGENVECTOR RESEARCH INCORPORATED

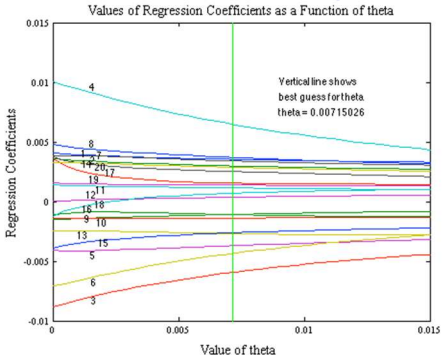
Ridge Regression

- Ridge Regression (RR) is one way to deal with ill-conditioned problems
- RR gets its name because a constant is added to the "ridge" of the covariance matrix in the formation of the pseudo-inverse:

$$\mathbf{X}^{+} = \left(\mathbf{X}^{T}\mathbf{X} + \mathbf{I}\theta\right)^{-1}\mathbf{X}^{T}$$

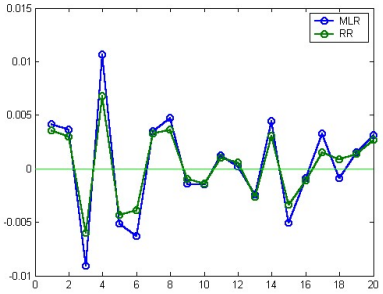
- The addition of the ridge ($\mathbf{I}\theta$ term) stabilizes the inverse and shrinks the values of the coefficients
  - this "ridging" is known as matrix regularization

RR Shrinkage




```
[brr,theta] = ridge(mx,my,0.015,31);
```


RR and MLR Regression Vectors



Problem with MLR and RR

- RR helps stabilize the inverse
  - ridging biases the regression
  - how to determine the ridge parameter θ?
- MLR does not work when $M<N_x$
- Possible solution: eliminate variables
  - how to choose which variables to keep?
    - stepwise regression or other variable selection
  - lose multivariate advantage - signal averaging
- Another solution: use PCA to reduce original variables to some smaller number of factors
  - retains multivariate advantage
  - noise reduction aspects of PCA

Principal Components Regression

- Principal Components Regression (PCR) is one way to deal with ill-conditioned problems
- Property of interest **y** is regressed on PCA scores:

$$\mathbf{X}^{+} = \mathbf{P}_K \left(\mathbf{T}_K^T \mathbf{T}_K \right)^{-1} \mathbf{T}_K^T$$

- Problem is to determine *K* the number of factors to retain in the formation of the model

Principal Components Regression

$$\mathbf{T}_K \mathbf{b}_{pc} = \mathbf{y} + \mathbf{e} = \mathbf{X}\mathbf{P}_K \mathbf{b}_{pc} \Leftrightarrow \mathbf{b} = \mathbf{P}_K \mathbf{b}_{pc}$$

$$\mathbf{T}_K \mathbf{b}_{pc} = \mathbf{y}$$

$$\mathbf{T}_K^T \mathbf{T}_K \mathbf{b}_{pc} = \mathbf{T}_K^T \mathbf{y}$$

$$\mathbf{b}_{pc} = \left(\mathbf{T}_K^T \mathbf{T}_k \right)^{-1} \mathbf{T}_K^T \mathbf{y}$$

$$\mathbf{b} = \mathbf{P}_K \left(\mathbf{T}_K^T \mathbf{T}_K \right)^{-1} \mathbf{T}_K^T \mathbf{y}$$

$$\mathbf{X}^{+} = \mathbf{P}_K \left(\mathbf{T}_K^T \mathbf{T}_K \right)^{-1} \mathbf{T}_K^T$$

- Note that $\mathbf{T}_K^T \mathbf{T}_K$ is *K*x*K* and square

Cross-Validation

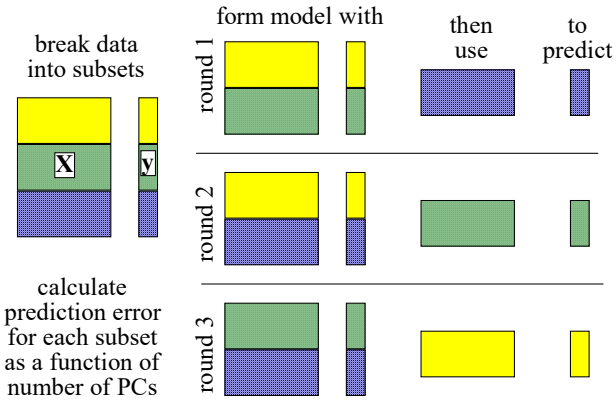
- Divide data set into *J* sample subsets
- For ***each subset*** (*j*=1,…, *J*):
  - Build PCA model using samples in the ***remaining*** subsets
  - Apply the model to subset *j* samples
  - Calculate PRESS (Predictive Residual Sum of Squares) for the subset samples:

$$\mathbf{e}_j^2 = \left(\mathbf{y} - \mathbf{X}\mathbf{b} \right)_j^2$$

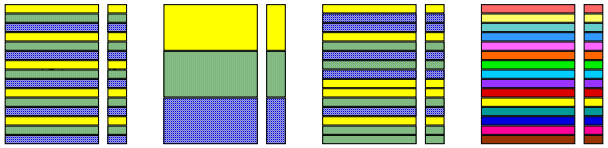
- Look for minimum or "knee" in CumPRESS curve

$$\text{RMSECV} = \left(\frac{1}{M} \sum_{j=1}^{J} \mathbf{e}_j^2 \right)^{1/2}$$

Cross-Validation Graphically



Formation of Test Sets



"Venetian blinds" - OK when data already in random order

contiguous blocks-best for time series

random selection-usually repeated several times

leave-one-out, used when not much data available

What else?

Custom selection, based on prior knowledge!

EIGENVECTOR RESEARCH INCORPORATED

Cross-Validation Considerations

- Cross-validation method selection criteria
  - *Number* of objects in dataset, *M*
  - *Order* of objects in dataset
  - *Objective* of cross-validation (specific type of error?)
  - Presence/absence of *replicates*
  - Remember the objective is to mimic future performance
- "Traps" to avoid
  - "Replicate sample trap"
    - Different replicates in both model and test set
  - "External subset selection trap" - extrapolation
    - Test set "space" outside of model set "space"

EIGENVECTOR RESEARCH INCORPORATED

Cross-validation Usage Matrix 1/2

DATASET TYPES	CROSS-VALIDATION METHODS				
	Venetian Blinds	Contiguous Blocks	Random Subsets	Leave-One Out	Custom
GENERAL CV Method Properties	• Easy • Relatively quick	• Easy • Relatively quick	• Easy • Can be slow, if m or number of iterations large • Selection of subsets unknown	• Easiest! (Only one parameter) • ***Avoid using if m>20***	• Flexible • Requires time to define splits
Small data sets (<~20 objects)	•	•	• OK, if many iterations done	• Good choice.... • ....unless DOE data	• often needed to avoid the ***external subset selection trap***
randomly-distributed objects	• Good choice	• Good choice	• Good choice • Can take a while with large m, many iterations	• OK, but.... • Can take a while with large m	•

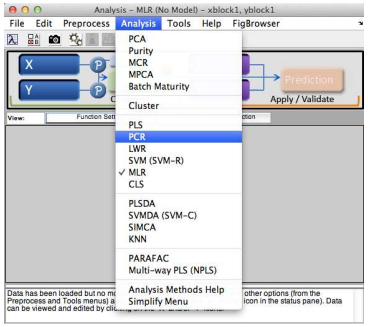
EIGENVECTOR RESEARCH INCORPORATED

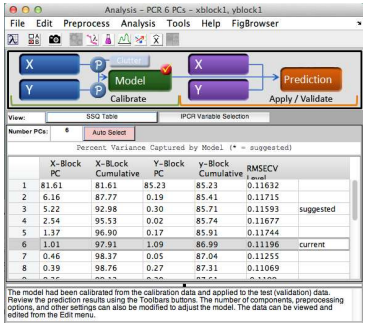
Cross-Validation Usage Matrix 2/2

DATASET TYPES	CROSS-VALIDATION METHODS				
	Venetian Blinds	Contiguous Blocks	Random Subsets	Leave-One Out	Custom
time-series data	• Useful for assessing ***NON-temporal*** model errors • Can be optimistic with low number of data splits	• Useful for assessing ***temporal*** stability of model	•	•	•
Batch data	• Useful for assessing predictability ***within*** batches/parts of batches	• Useful for assessing predictability ***between*** batches/parts of batches	•	•	• Can manually select "batch-wise" test sets
Blocked data (replicates)	• Beware the ***replicate sample trap*** (optimistic results)!	• Good way to avoid ***replicate sample trap*** • Beware the ***external subset selection trap***!	• Can use to avoid ***replicate sample trap*** (high number of splits, high iterations preferable)	• overly optimistic results, due to ***replicate sample trap***	•
Designed Experiment (DOE) data	• Dangerous, unless object order is randomized	• Dangerous, unless object order is randomized	•	• Not recommended (***external subset selection trap***)	• often needed to avoid the ***external subset selection*** *trap*

EIGENVECTOR RESEARCH INCORPORATED

Switch Analysis from MLR to PCR, Calculate Model



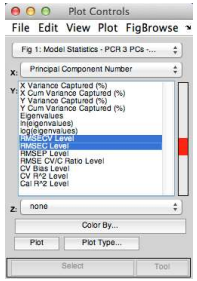


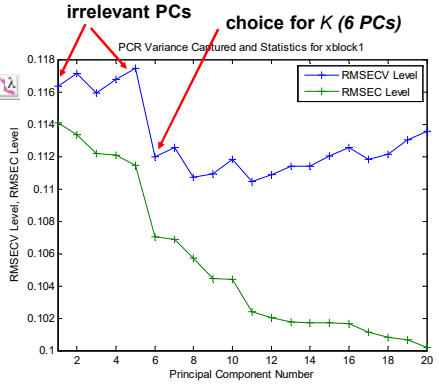
EIGENVECTOR RESEARCH INCORPORATED

PCR Cross-Validation Example

Recall: time series data so contiguous block CV

Click variance "lambda" icon to get CV results





EIGENVECTOR RESEARCH INCORPORATED

PCR Variance Captured

Percent Variance Captured by PCR Model

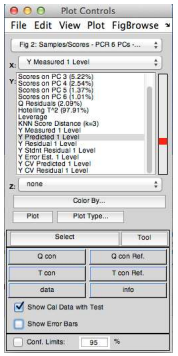
PC #	X-Block This PC	X-Block Total	Y-Block This PC	Y-Block Total
1	81.61	81.61	85.23	85.23
2	6.16	87.77	0.19	85.41
3	5.22	92.98	0.30	85.71
4	2.54	95.53	0.02	85.74
5	1.37	96.90	0.17	85.91
6	1.01	97.91	1.09	86.99
7	0.46	98.37	0.05	87.04
8	0.39	98.76	0.27	87.31
9	0.36	99.12	0.30	87.61
10	0.24	99.37	0.02	87.63

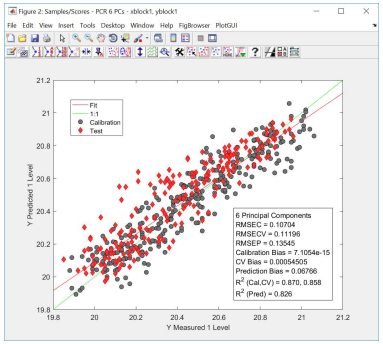
EIGENVECTOR RESEARCH INCORPORATED

PCR Model Fit to Calibration Data and Validation Predictions

Click scores "flask" to get prediction results

Adjust Plot Controls to get desired plot





EIGENVECTOR RESEARCH INCORPORATED

Problems with PCR

- Some PCs not relevant for prediction, but are only relevant for describing variance in **X**
  - leads to local minima and increase in PRESS
- This is a result of PCs determined without using information about property to be predicted **y**
- A solution is to find factors using information from **y** and **X**

EIGENVECTOR RESEARCH INCORPORATED

Partial Least Squares

- PLS is related to PCR and MLR
  - PCR captures maximum variance in **X**
  - MLR achieves maximum correlation between **X** and **Y**
  - PLS tries to do both by maximizing covariance between **X** and **Y**
- Requires addition of weights **W** to maintain orthogonal scores
- Factors calculated sequentially by projecting **Y** through **X**

$$\mathbf{X}^{+} = \mathbf{R}_K\left(\mathbf{T}_K^{\mathrm{T}}\mathbf{T}_K\right)^{-1}\mathbf{T}_K^{\mathrm{T}} = \mathbf{W}_K\left(\mathbf{P}_K^{\mathrm{T}}\mathbf{W}_K\right)^{-1}\left(\mathbf{T}_K^{\mathrm{T}}\mathbf{T}_K\right)^{-1}\mathbf{T}_K^{\mathrm{T}}$$

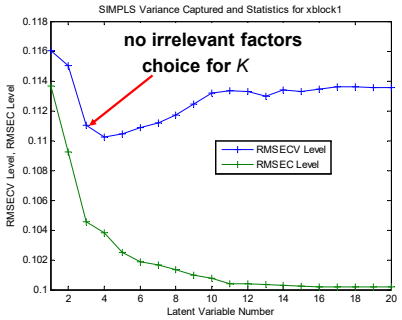
EIGENVECTOR RESEARCH INCORPORATED

PLS Cross-Validation Example

Set Analysis to PLS

Calculate model

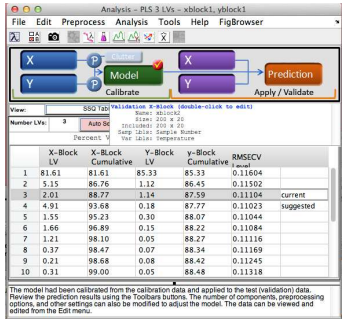
Click variance "lambda" to get CV results 

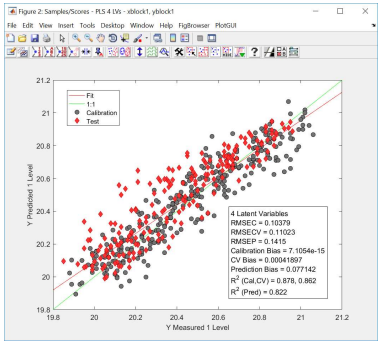


EIGENVECTOR RESEARCH INCORPORATED

PLS Model Fit to Calibration Data and Validation Predictions

Set number of LVs to 4





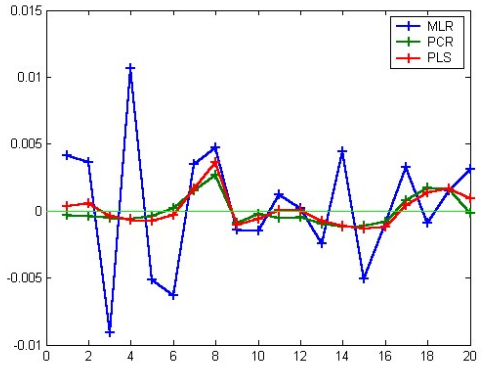
EIGENVECTOR RESEARCH INCORPORATED

PLS Variance Captured

Percent Variance Captured by PLS Model

LV #	X-Block This LV	X-Block Total	Y-Block This LV	Y-Block Total
1	81.61	81.61	85.33	85.33
2	5.15	86.76	1.12	86.45
3	2.01	88.77	1.14	87.59
4	4.91	93.68	0.18	87.77
5	1.55	95.23	0.30	88.07
6	1.66	96.89	0.15	88.22
7	1.21	98.10	0.05	88.27
8	0.37	98.47	0.07	88.34
9	0.21	98.68	0.08	88.42
10	0.31	99.00	0.05	88.48

Regression Vectors



NIPALS: PLS Algorithm

Choose $\mathbf{u}_1 = \mathbf{y}$ or one column of $\mathbf{Y}$

$$\mathbf{w}_1 = \frac{\mathbf{X}^T\mathbf{u}_1}{\|\mathbf{X}^T\mathbf{u}_1\|} \quad (1)$$

$$\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1 \quad (2)$$

$$\mathbf{q}_1 = \frac{\mathbf{u}_1^T\mathbf{t}_1}{\|\mathbf{u}_1^T\mathbf{t}_1\|} \quad (3)$$

$$\mathbf{u}_1 = \mathbf{Y}\mathbf{q}_1 \quad (4)$$

Check for convergence by comparing $\mathbf{t}_1$ to previous $\mathbf{t}_1$. If $\mathbf{Y} = \mathbf{y}$ skip (3) and (4) and continue

$$\mathbf{p}_1 = \frac{\mathbf{X}^T\mathbf{t}_1}{\|\mathbf{t}_1^T\mathbf{t}_1\|} \quad (5)$$

$$\mathbf{p}_{1new} = \frac{\mathbf{p}_{1old}}{\|\mathbf{p}_{1old}\|} \quad (6)$$

$$\mathbf{t}_{1new} = \mathbf{t}_{1old}\|\mathbf{p}_{1old}\| \quad (7)$$

$$\mathbf{w}_{1new} = \mathbf{w}_{1old}\|\mathbf{p}_{1old}\| \quad (8)$$

Find the regression coefficient for the inner relation:

$$b_1 = \frac{\mathbf{u}_1^T\mathbf{t}_1}{\mathbf{t}_1^T\mathbf{t}_1} \quad (9)$$

After calculating scores and loadings for first Latent Variable, the X and Y-block residuals are calculated:

$$\mathbf{E}_1 = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1^T \quad (10)$$

$$\mathbf{F}_1 = \mathbf{Y} - \mathbf{u}_1\mathbf{q}_1^T \quad (11)$$

Repeat entire procedure replacing $\mathbf{X}$ and $\mathbf{Y}$ with their residuals

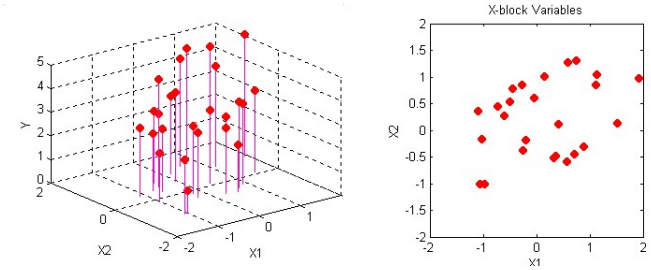
Non-linear iterative patial least squares (NIPALS).
Geladi, Paul; Kowalski, Bruce (1986), "Partial Least Squares Regression:A Tutorial", *Analytica Chimica Acta* **185**: 1–17, doi:10.1016/0003-2670(86)80028-9

Other PLS Algorithms

- It can be shown that $\mathbf{w}_1$ is given by
$$\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w}_1 = \lambda\mathbf{w}_1$$
- The SIMPLS algorithm uses an orthogonalization of a Krylof sequence (faster than NIPLS algorithm)
- The important thing to remember is:
  PLS finds factors in **X** which are correlated with **Y** while describing large amounts of variance in **X**

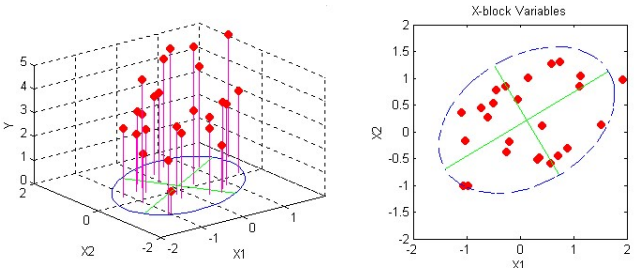
Sijmen de Jong, "SIMPLS: an alternative approach squares regression," *Chemometrics and Intelligent Laboratory Systems*, **18** (1993) 251-263

Y Projected onto X Plane



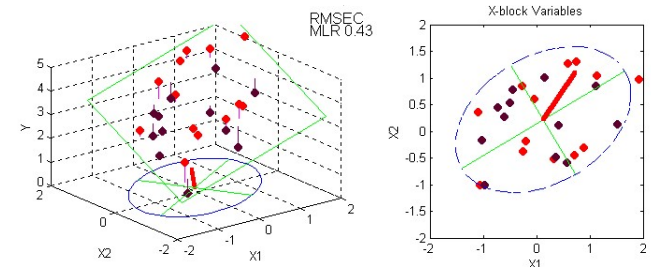
EIGENVECTOR RESEARCH INCORPORATED

PCA of X-Block



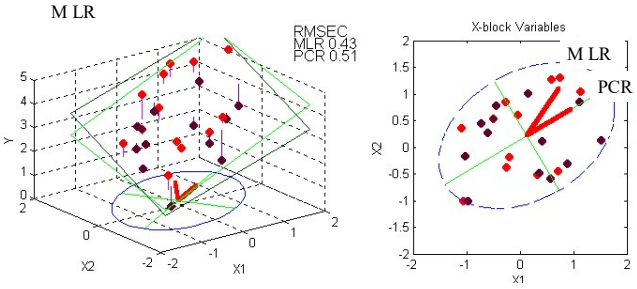
EIGENVECTOR RESEARCH INCORPORATED

MLR Regression Vector and Surface



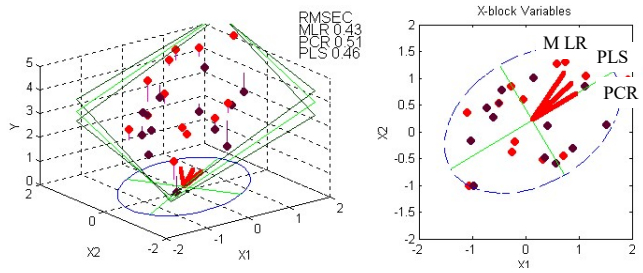
EIGENVECTOR RESEARCH INCORPORATED

PCR Regression Vector and Surface



EIGENVECTOR RESEARCH INCORPORATED

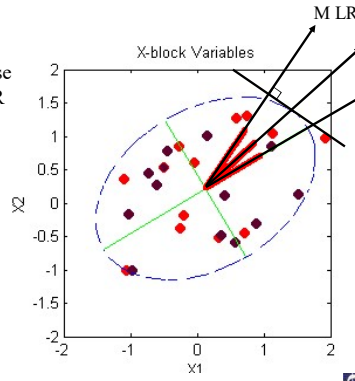
PLS Regression Vector and Surface



EIGENVECTOR RESEARCH INCORPORATED

Geometric Relationship of MLR, PCR, and PLS

PLS is the vector on the PCR ellipse upon which MLR has the longest projection



Line perpindicular to the MLR regression vector

EIGENVECTOR RESEARCH INCORPORATED

PLS for Multivariate Y

- PLS can be used to relate multivariate **X** to multivariate **Y** (*a.k.a.*, PLS2)
  - outer relationships

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

$$\mathbf{Y} = \mathbf{U}_K\mathbf{Q}_K^T + \mathbf{F}$$

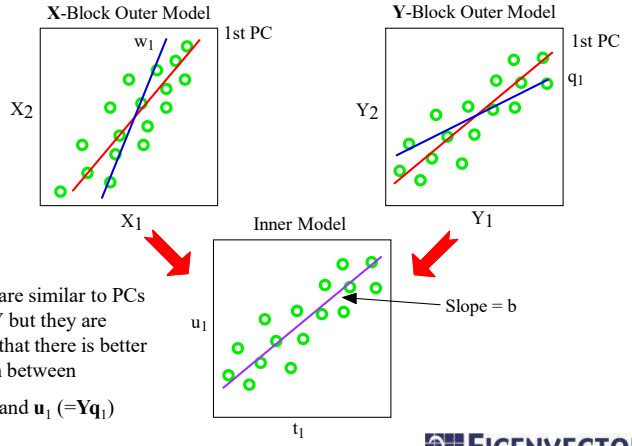
  - inner relationship

$$\mathbf{U}_K = \mathbf{T}_K\mathbf{B}_K$$

  - *i.e.,* the scores in **Y** are linear combinations of the scores in **X**

EIGENVECTOR RESEARCH INCORPORATED

PLS2



$\mathbf{w}_1$ and $\mathbf{q}_1$ are similar to PCs in **X** and **Y** but they are rotated so that there is better correlation between

$\mathbf{t}_1$ (=$\mathbf{Xw}_1$) and $\mathbf{u}_1$ (=$\mathbf{Yq}_1$)

EIGENVECTOR RESEARCH INCORPORATED

Model Quality Measures

- Root Mean Square Error (RMSE) Metrics
  - RMSE**C**
  - RMSE**CV**
  - RMSE**P**
  - In units of the **Y** variable!
- Correlation Coefficient (r)
  - Unit-less
  - Considers the range of **Y**

$$\sqrt{\frac{\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2}{m}}$$

$$\frac{\sum_{i=1}^{m}\left(\hat{y}_i-\hat{\bar{y}}\right)\left(y_i-\bar{y}\right)}{\sqrt{\left[\sum_{i=1}^{m}\left(\hat{y}_i-\hat{\bar{y}}\right)\sum_{i=1}^{m}\left(y_i-\bar{y}\right)\right]}}$$

Root Mean Square Error (RMSE) Metrics

- These are used to assess a model's *fit to the data* and *predictive ability on new data*
- Measures "average" deviation of model estimates from the measured data
- Measure of *fit* - root mean squared error of ***calibration*** (RMSE**C**)

$$\text{RMSE}\mathbf{C}=\sqrt{\frac{\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2}{m}}$$

*i's refer to all samples **used to build the model***

Cross-Validation Error

- RMSE**C** measures *fit to the model data*. RMSECV (root mean squared error of cross-validation) is an estimate of *predictive power on new data*.
- RMSE**CV** is a function of the number of factors *k and* how the test sets were selected

$$RMSECV=\sqrt{\frac{\sum_{j=1}^{J}\sum_{i=1}^{m_j}\left(y_i-\hat{y}_i\right)^2}{m_J}}=\sqrt{\frac{PRESS}{m_J}}$$

J's refer to different CV subsets

*i's refer to CV subset samples- **not used to build CV models***

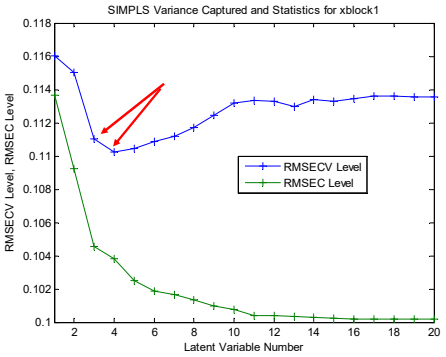
Prediction Error

- Prediction error is often used to validate a model and is a true measure of the *predictive power on new data*
- Measure of *prediction error* - root mean squared error prediction (RMSE**P**)

$$\text{RMSE}\mathbf{P}=\sqrt{\frac{\sum_{i=1}^{m}\left(y_i-\hat{y}_i\right)^2}{m}}$$

*i's refer to samples **NOT** used to build the model*

RMSE metrics, as a function of factor (PC, LV)



RMSEC and RMSECV can also be used to determine the optimal number of factors (LVs, PCs) to be used in a model

EIGENVECTOR RESEARCH INCORPORATED

Comparison of Models

- MLR, PCR, and PLS models were constructed using SFCM data: Calibration used (xblock1) and test used (xblock2).

	MLR	PCR	PLS
RMSEC	0.1002	0.1070	0.1038
RMSECV	0.1136	0.1120	0.1102
RMSEP	0.1498	0.1355	0.1415

- ***Fit*** and ***prediction*** are two entirely different aspects of a model's performance

EIGENVECTOR RESEARCH INCORPORATED

Number of PCs or LVs

- Choice is not always simple
- A few rules of thumb
  - sqrt(*M*) a good choice for number of splits
  - useful to do repeated CVs with different data ordering
  - if data is time series use block CV due to correlated noise
  - be conservative, models are more often over-fit than under-fit
  - best choice is often not the global minimum PRESS
  - look for minimum of PRESS and work backwards if improvement is not at least 2%
  - RMSEC<RMSECV by more than ~20% indicates overfit
  - look at variance captured in **X** and **Y**. Is it significant with respect to what you know about the data?

EIGENVECTOR RESEARCH INCORPORATED

Model Diagnostics

- Diagnostics useful for finding outliers/uniques
- **X**-block Q residual and T^2
- **X**-block leverage and studentized **Y**-block residuals
- Try SFCM example ***without removing outliers***

EIGENVECTOR RESEARCH INCORPORATED

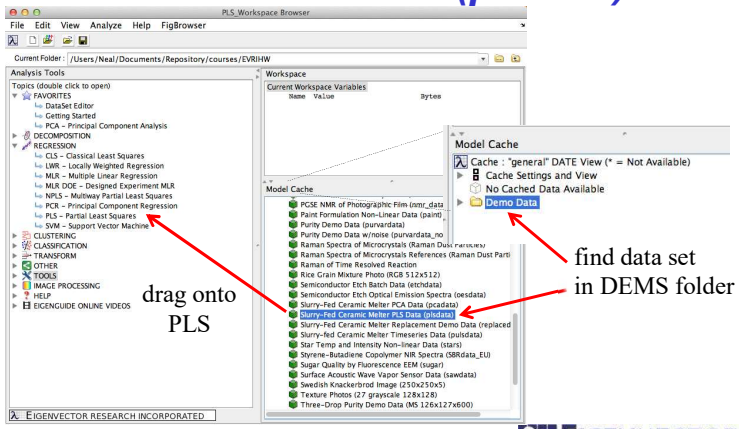
Build PLS Model on SFCM Data

- Construct a linear regression for yblock1 from xblock1 (time series data)
  - predict level of slurry fed ceramic melter (Y-block)
  - using melter temperatures (X-block)
- Test the model on xblock2 and yblock2

IF data still loaded, can do:
- Edit/Calibration/X-block Data
- Row Labels Tab
- right-click on "Incl" : "Clear/Reset" (checks all rows)
- calculate model

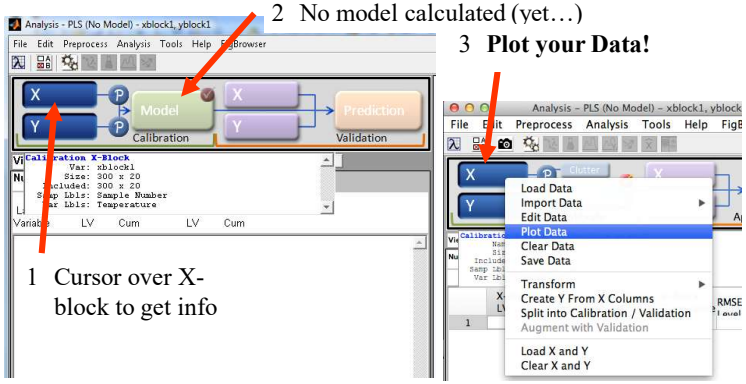
EIGENVECTOR RESEARCH INCORPORATED

Load SFCM Data (plsdata)



EIGENVECTOR RESEARCH INCORPORATED

Data: loaded but not analyzed

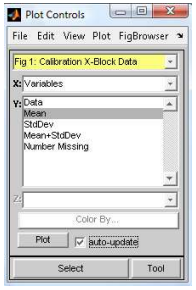


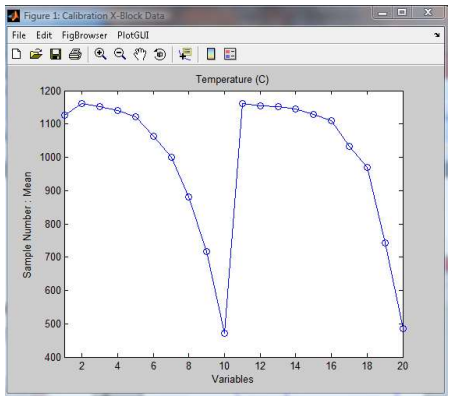
EIGENVECTOR RESEARCH INCORPORATED

Plot Your Data

Default plot

Column / Variable mean





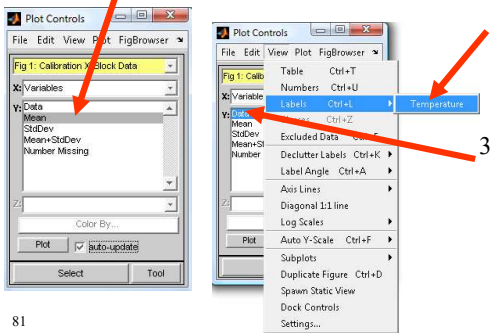
EIGENVECTOR RESEARCH INCORPORATED

Plot Your Data

1 Plot control default

can look at summary stats

the Plot control generates plots in MATLAB figure windows



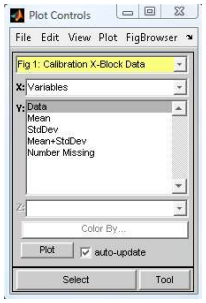
2 under view menu check labels:Temperature

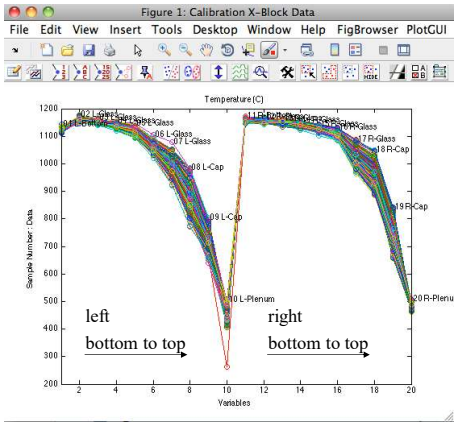
3 under Y: menu highlight Data

EIGENVECTOR RESEARCH INCORPORATED

Plot Your Data

all samples vs. variables



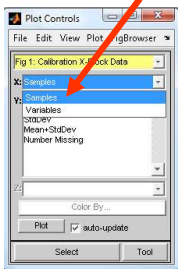


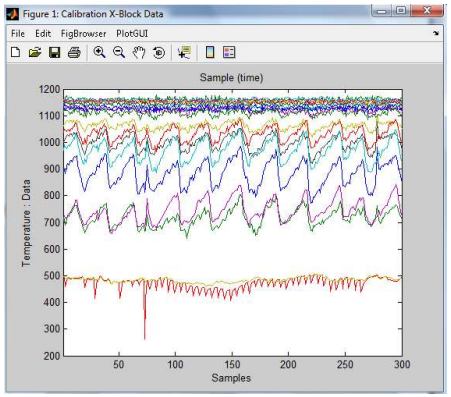
EIGENVECTOR RESEARCH INCORPORATED

Plot Your Data

all variables vs. samples (time)

under X: menu highlight Samples

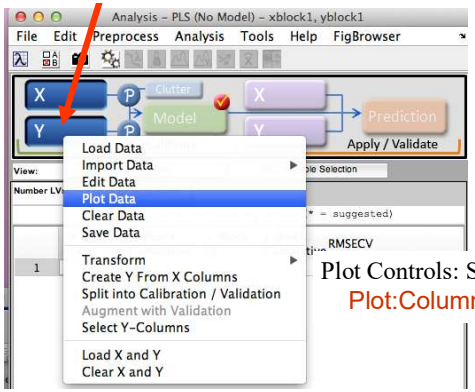




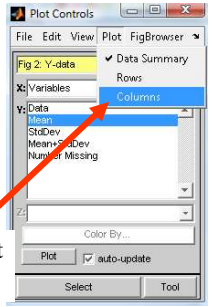
EIGENVECTOR RESEARCH INCORPORATED

Plot Y data

Click

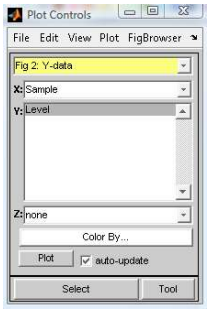


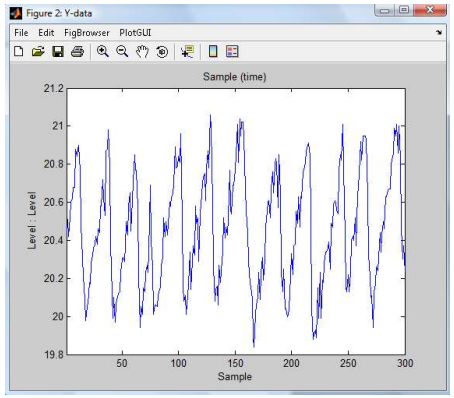
Plot Controls: Select Plot:Columns



EIGENVECTOR RESEARCH INCORPORATED

Plot Y Data





EIGENVECTOR RESEARCH INCORPORATED

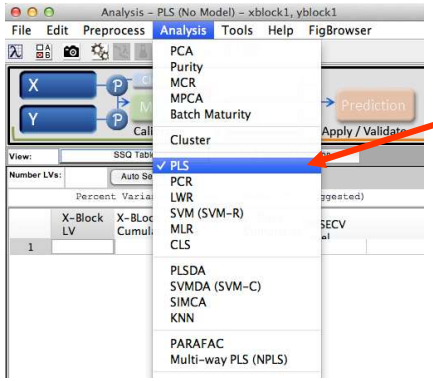
Plot Your Data: Summary

- Bottom temperatures higher than top temperatures
  - surface and plenum space is cooler than the bottom
- Trend in time
  - "saw-tooth" pattern showing correlation between some temperatures and level

EIGENVECTOR RESEARCH INCORPORATED

Which Regression?

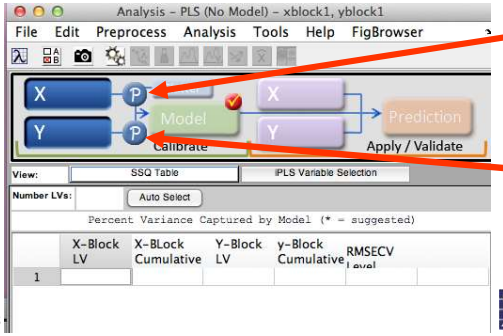
- BACK to Analysis Window, then Analysis menu
- Choose PLS...



EIGENVECTOR RESEARCH INCORPORATED

How Should We Scale the Data?

- Variables are in same units and there's and reason to believe that variance is associated with signal: Suggests mean-centering.
- X Preprocessing is set under Preprocess:X-block

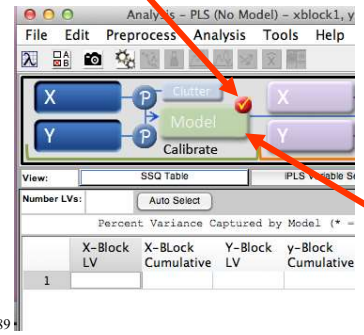


EIGENVECTOR RESEARCH INCORPORATED

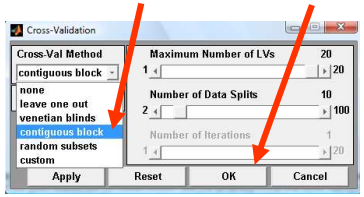
How to Cross-Validate?

- Time series data suggests contiguous block cross-validation

1 click check for Cross-Validation



2 contiguous block, then OK

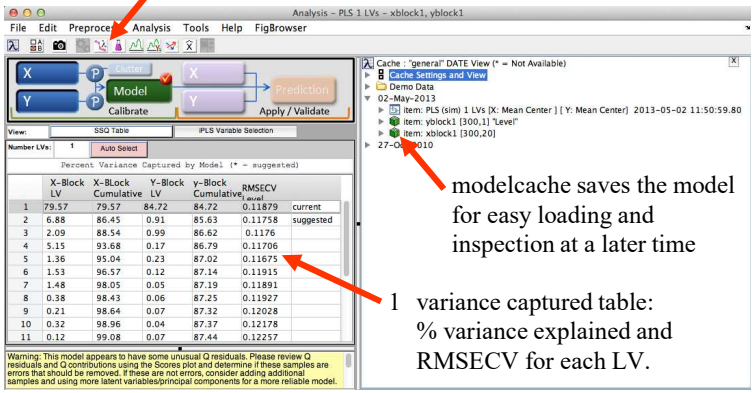


3 click Model button to perform cross-validation *and* build the regression model

EIGENVECTOR RESEARCH INCORPORATED

Regression Results

2 Click Eigenvalue button to plot RMSEC and RMSECV

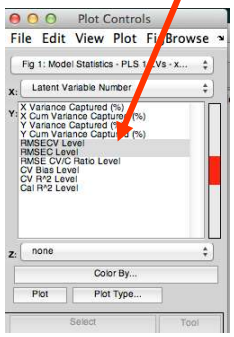


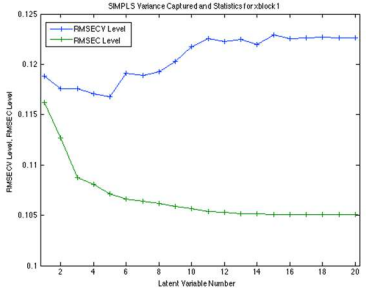
modelcache saves the model for easy loading and inspection at a later time

1 variance captured table: % variance explained and RMSECV for each LV.

RMSECV Plot

Plot RMSEC and RMSECV vs. LV.

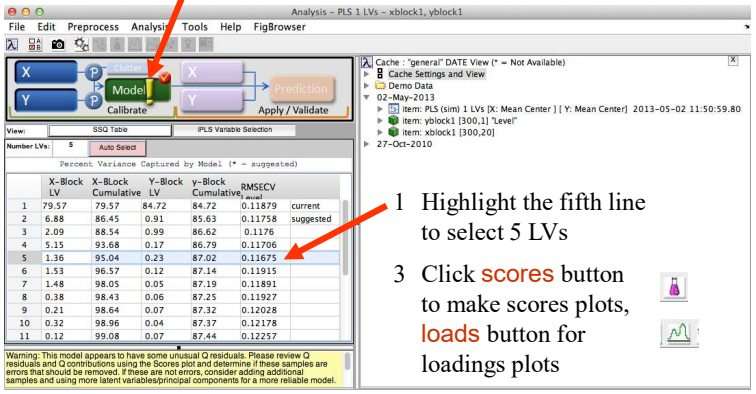




EIGENVECTOR RESEARCH INCORPORATED

Choose Number of LVs

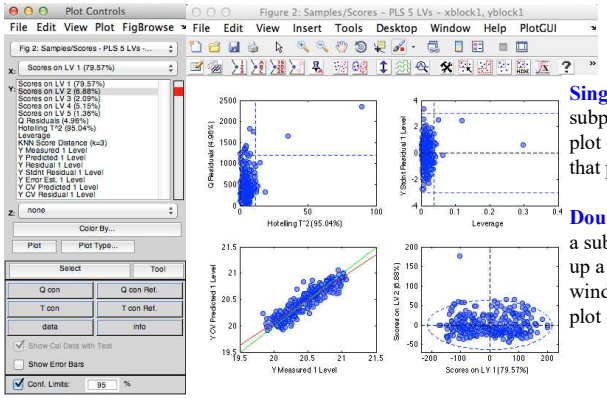
2 Click Model button to reconstruct the model with 5 LVs



1 Highlight the fifth line to select 5 LVs

3 Click scores button to make scores plots, loads button for loadings plots

Scores Summary Plot



Single-click on a subplot: brings up plot controls for that plot

Double-click on a subplot: brings up a new figure window with that plot only

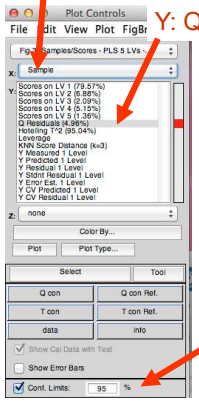
EIGENVECTOR RESEARCH INCORPORATED

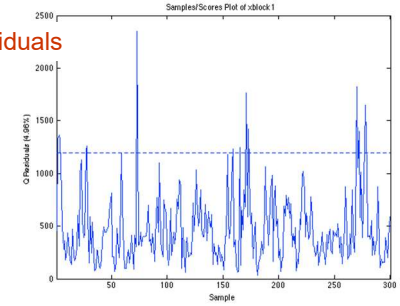
X-Block Q Residuals (by sample)

1 set plot controls

X: Sample

Y: Q Residuals





2 Check Conf. Limits

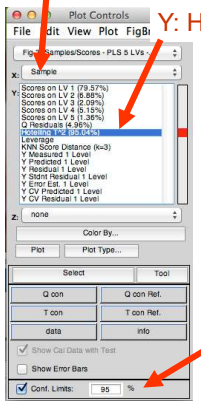
EIGENVECTOR RESEARCH INCORPORATED

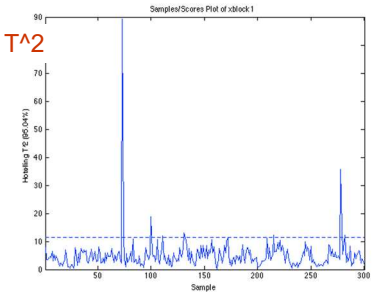
X-Block Hotelling T^2 (by sample)

1 set plot controls

X: Sample

Y: Hotelling T^2





2 Check Conf. Limits

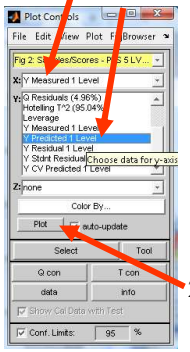
EIGENVECTOR RESEARCH INCORPORATED

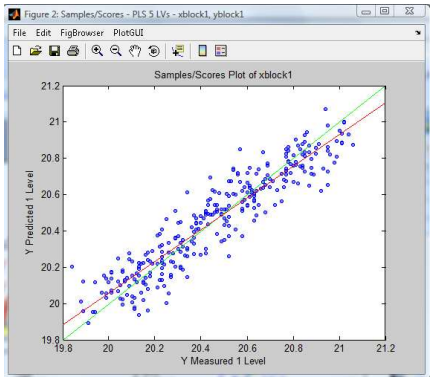
Calibration Curve (Predicted vs. Measured)

1 Click Scores Button

X:Y Measured

Y:Y Predicted





2 Click Plot if auto-update is not checked

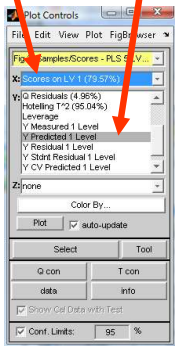
EIGENVECTOR RESEARCH INCORPORATED

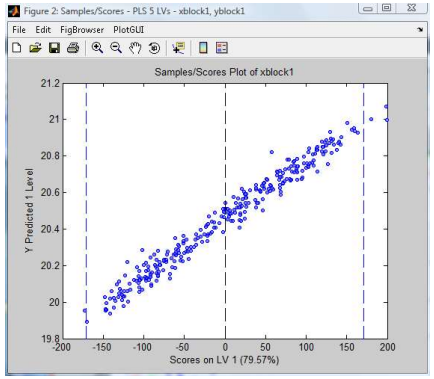
Predicted Y vs. LV 1

1 Click Scores Button

X:LV 1

Y:Y Predicted





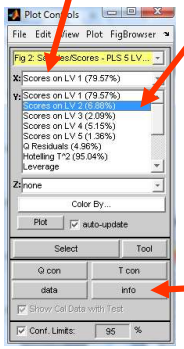
EIGENVECTOR RESEARCH INCORPORATED

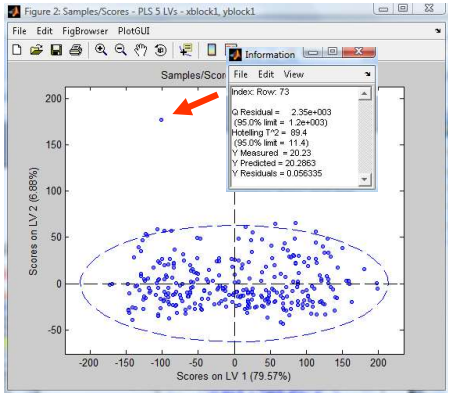
LV 2 vs. LV 1

1 Click Scores Button

X:Scores on LV 1

Y:Scores on LV 2





2 Click info Button and select sample

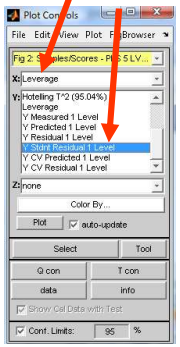
EIGENVECTOR RESEARCH INCORPORATED

Studentized Y Residual vs. Leverage

1 Click Scores Button

X:Leverage

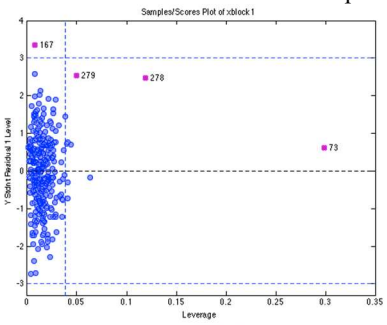
Y:Y Stdnt Residual



2 select outliers

3 view numbers

samples removed in first example



EIGENVECTOR RESEARCH INCORPORATED

Calculation of Studentized Residuals

- Given the pseudo-inverse $\mathbf{X}^+$ the leverage for a sample $\mathbf{x}_m$ = x(m,:) and column $\mathbf{X}^+(:,m)$ is

$$l_m = \mathbf{x}_m \mathbf{X}^+(:,m)$$

- Studentized residuals for column m^{th} of $\mathbf{y}$, $t_{e,m}$

$$e_m = \hat{y}_m - y_m$$

$$\sigma = \left(\frac{1}{M-K}\mathbf{e}^T\mathbf{e}\right)^{1/2}$$

$$t_{e,m} = \frac{e_m}{\sigma(1-l_m)^{1/2}}$$

A studentized residual is the quotient resulting from the division of a residual by an estimate of its standard deviation and is a form of t-statistic.

EIGENVECTOR RESEARCH INCORPORATED

How Much Leverage is Too Much?

- In PLS and PCR a good rule of thumb is $3K/M$, where K is the number of LVs or PCs, and M is the number of samples
- In MLR, use $2N_x/M$, where N_x is the number of X-block variables

EIGENVECTOR RESEARCH INCORPORATED

Regression Example


1,2-


cis-


trans-


styrene

- NIR transmission spectra of styrene-butadiene copolymers
- Different amounts of 4 analytes
  - **All 4** are known for **all** samples (by NMR)
- Data file: SBRdata_EU.mat
  - 60 calibration samples in arrays Xcal, Ycal
  - 10 test samples in arrays Xtest, Ytest

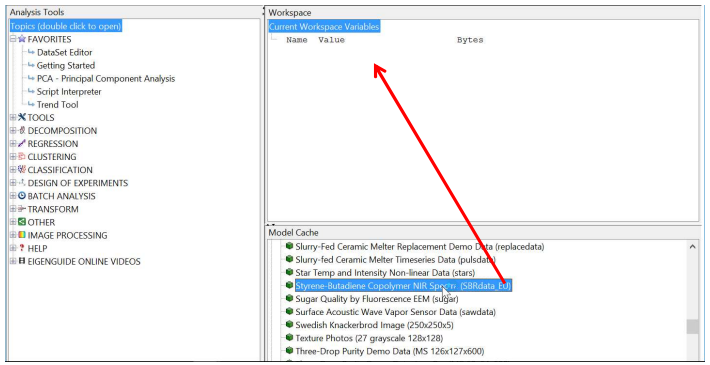
EIGENVECTOR RESEARCH INCORPORATED

Regression Methods

- Compare the styrene predictions using the four different regression approaches below.
- CLS – no centering
- MLR (stepwise) – no centering
- PCR – no centering
- PLS – no centering
- Additionally: show results with mean centering

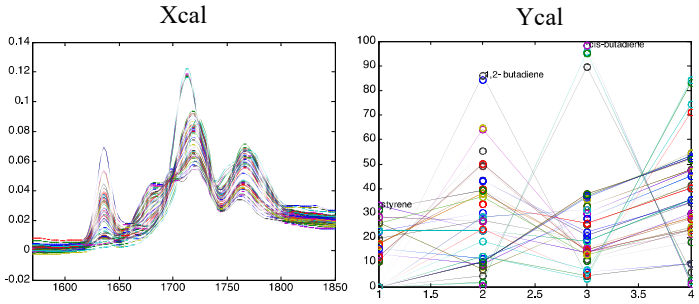
EIGENVECTOR RESEARCH INCORPORATED

Load Into Workspace



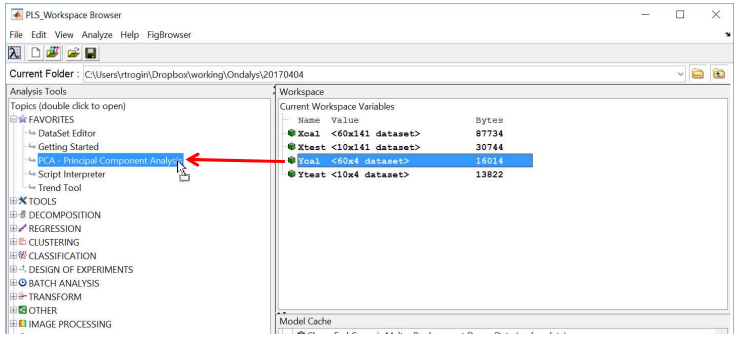
EIGENVECTOR RESEARCH INCORPORATED

Load and Plot Calibration Data



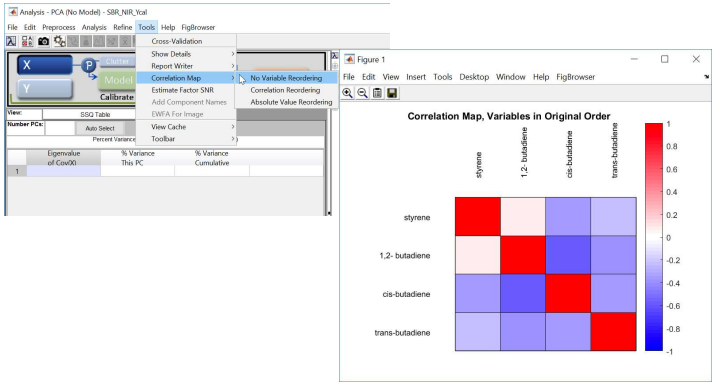
EIGENVECTOR RESEARCH INCORPORATED

Try This:



EIGENVECTOR RESEARCH INCORPORATED

Correlation Map



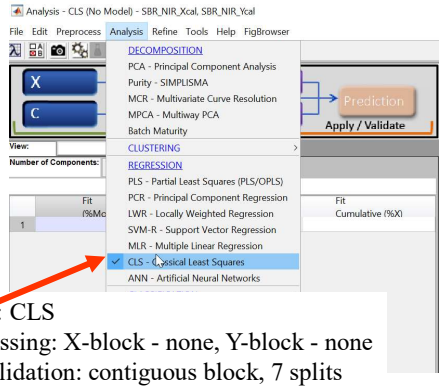
EIGENVECTOR RESEARCH INCORPORATED

Loading Data

- Calibration data
  - Xcal as x-block
  - Ycal as y-block
- Validation data
  - Xtest as x-block
  - Ytest as y-block

EIGENVECTOR RESEARCH INCORPORATED

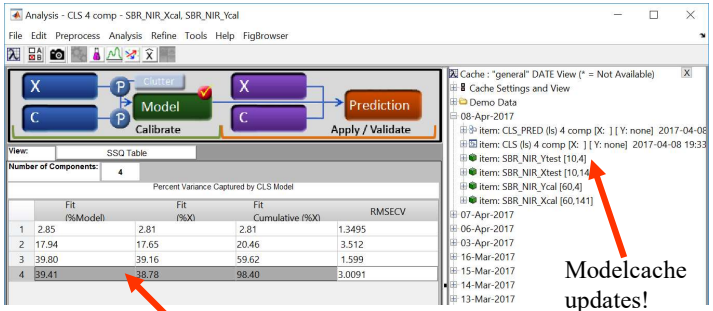
CLS Regression



1 Analysis: CLS
2 Preprocessing: X-block - none, Y-block - none
3 Cross-validation: contiguous block, 7 splits
4 Calculate model

EIGENVECTOR RESEARCH INCORPORATED

CLS results



Modelcache updates!

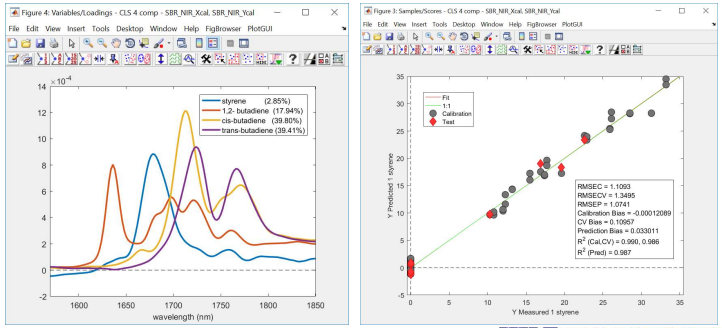
4 components determined "automatically" (because of 4 Y variables)

EIGENVECTOR RESEARCH INCORPORATED

CLS Results

Loadings: **all 4 components**

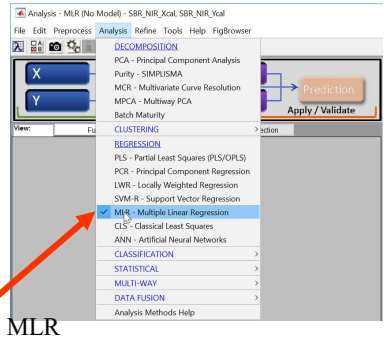
Scores: Styrene predicted vs. Styrene measured, Cal data included



This is the **S** matrix!

EIGENVECTOR RESEARCH INCORPORATED

MLR- styrene

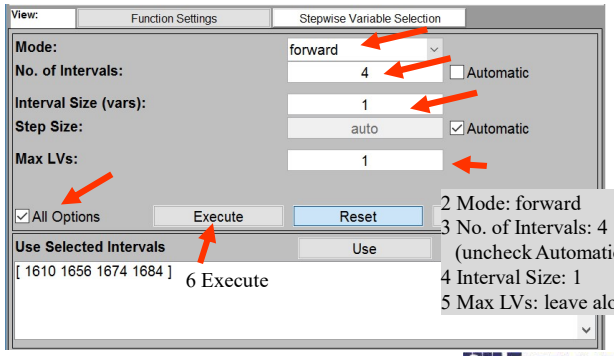


1 Analysis: MLR
2 Preprocessing: X-block - none, C-block - none
3 Cross-validation: contiguous block, 7 splits
4 Right-click calibration Y-block =>Select Y Columns => Styrene
5 Stepwise Variable Selection button

EIGENVECTOR RESEARCH INCORPORATED

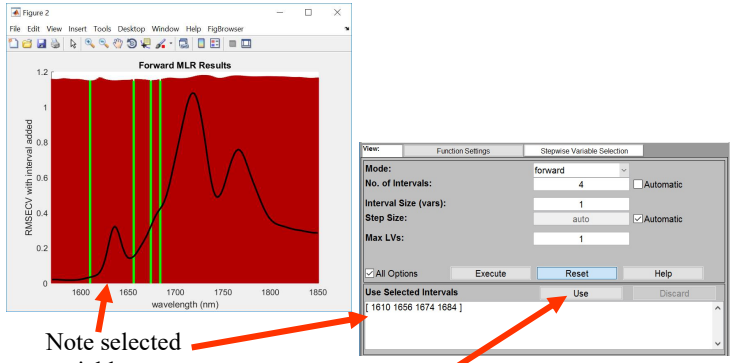
Stepwise Variable Selection

1 Check "All Options"
6 Execute



EIGENVECTOR RESEARCH INCORPORATED

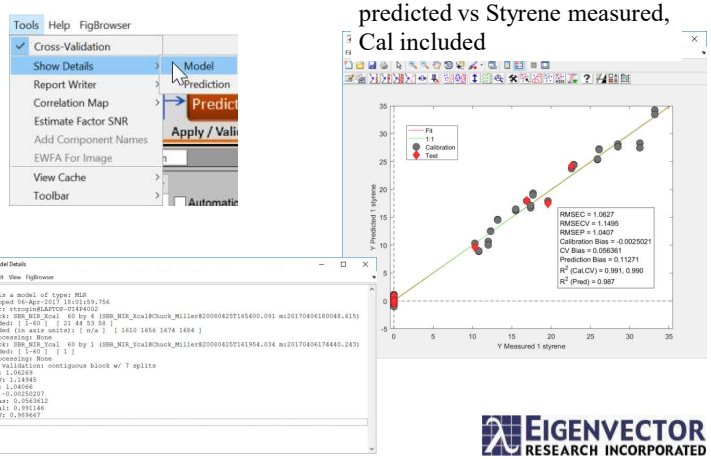
Forward MLR Results



EIGENVECTOR RESEARCH INCORPORATED

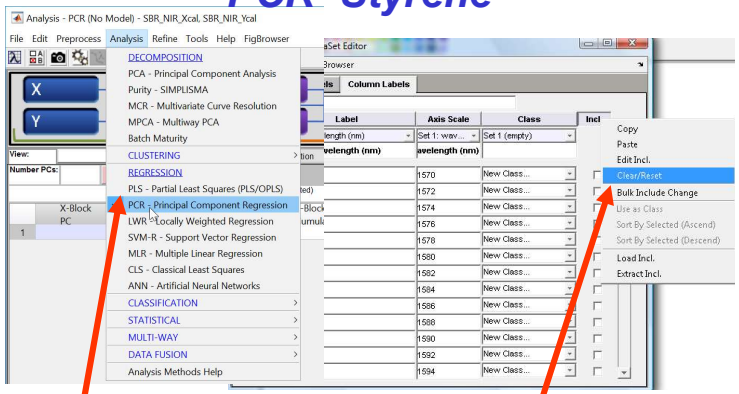
MLR results

Scores button: Styrene predicted vs Styrene measured, Cal included



EIGENVECTOR RESEARCH INCORPORATED

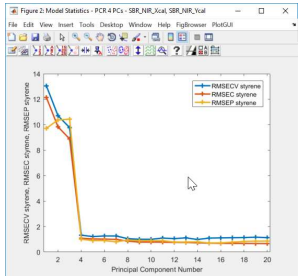
PCR- Styrene



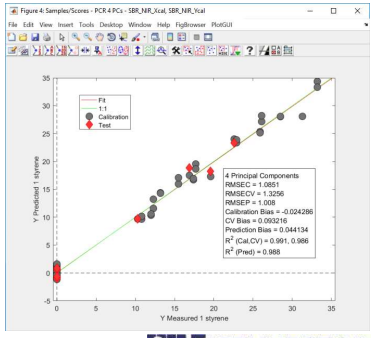
EIGENVECTOR RESEARCH INCORPORATED

PCR Results

Eigenvalue button:
RMSE[C], [CV], and [P]



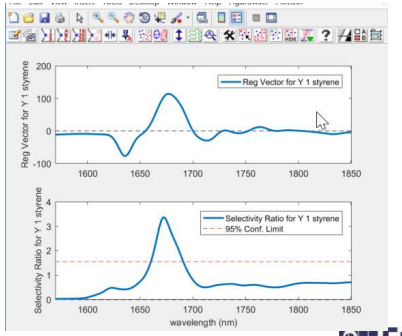
Scores: Styrene predicted vs.
Styrene measured



EIGENVECTOR RESEARCH INCORPORATED

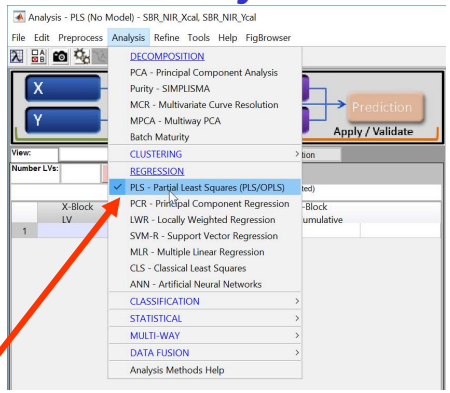
PCR Results

Loadings button: Regression Vector for Styrene
Type "2" (creates second plot)
Second plot: Selectivity Ratio for Styrene



EIGENVECTOR RESEARCH INCORPORATED

PLS Styrene

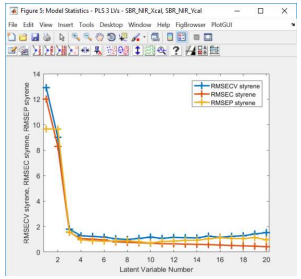


1 Analysis: PLS
2 Calculate

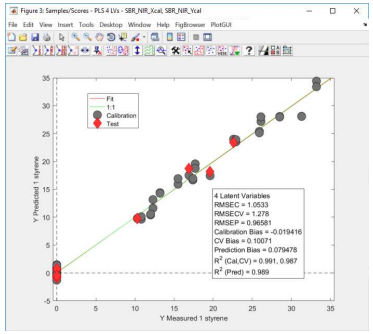
EIGENVECTOR RESEARCH INCORPORATED

PLS Results

Eigenvalue button:
RMSE[C], [CV], and [P]

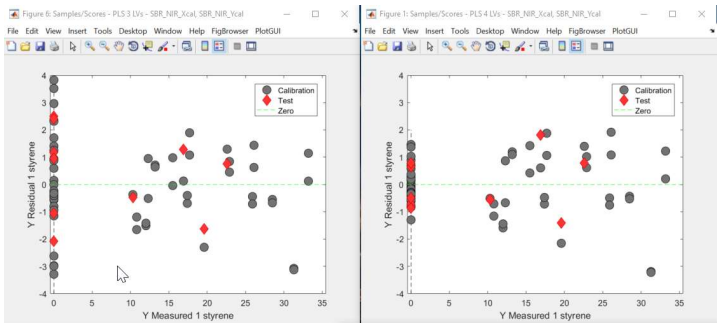


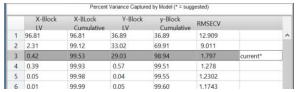
Scores button: Predicted vs.
Measured Styrene



EIGENVECTOR RESEARCH INCORPORATED

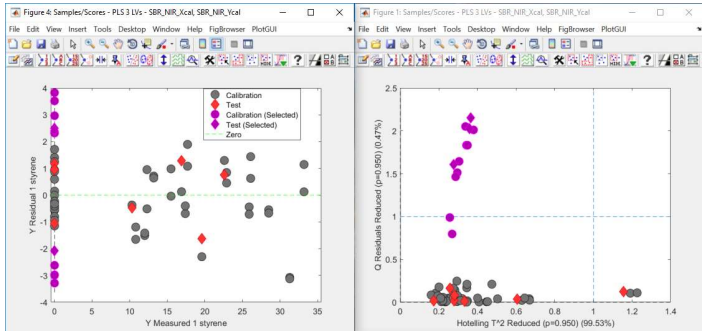
PLS – 3 vs 4 LVs





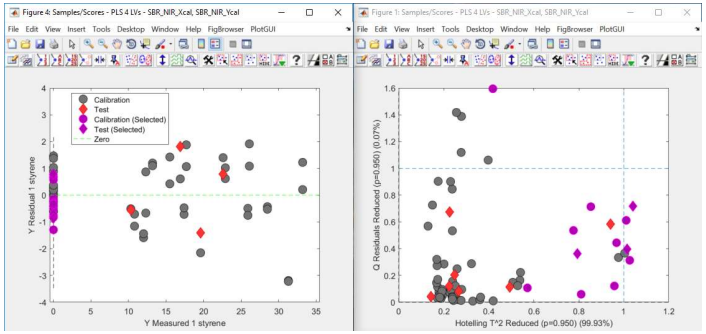
EIGENVECTOR RESEARCH INCORPORATED

Back to 3 LVs



EIGENVECTOR RESEARCH INCORPORATED

And at 4 LVs



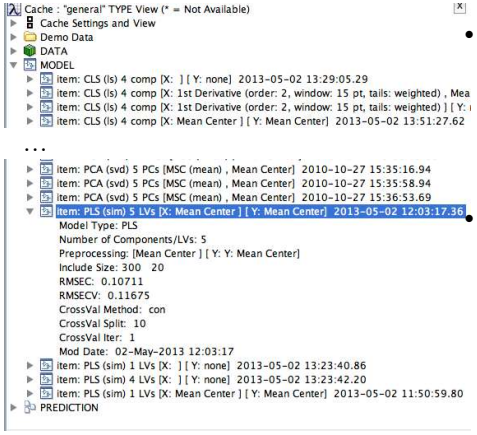
EIGENVECTOR RESEARCH INCORPORATED

Regression Summary- Styrene

	RMSEC	CV	P	Comments
Not centered				
CLS	1.11	1.35	1.07	4 factors
MLR	1.06	1.15	1.04	4 stepwise-selected variables
PCR	1.08	1.33	1.01	4 factors
PLS	1.05	1.28	0.97	4 factors
Mean centered				
CLS	1.39	1.61	1.33	4 factors
MLR	0.89	1.07	0.97	4 stepwise-selected variables
PCR	0.84	0.94	0.73	4 factors
PLS	0.84	0.95	0.73	4 factors

EIGENVECTOR RESEARCH INCORPORATED

Modelcache

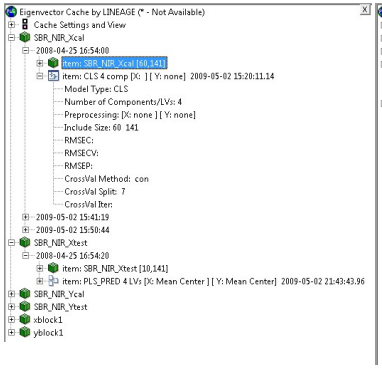


- Can list recent work in 3 ways:
  - By type (shown here)
  - By lineage
  - By date
- Also, can load/show/save any item in the cache

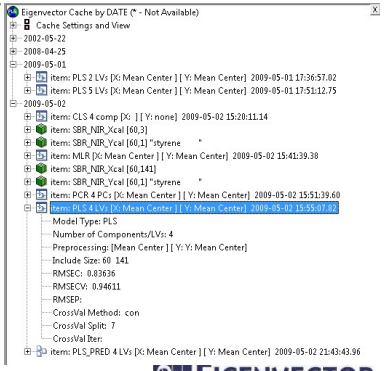
EIGENVECTOR RESEARCH INCORPORATED

Modelcache by lineage and date

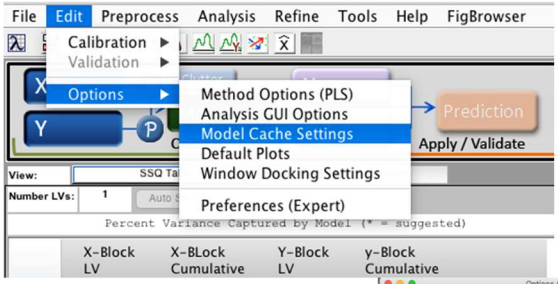
By lineage

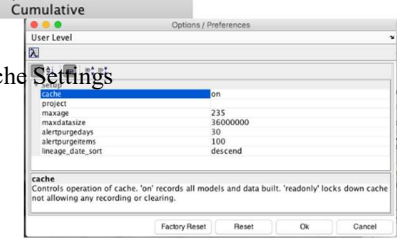


By date



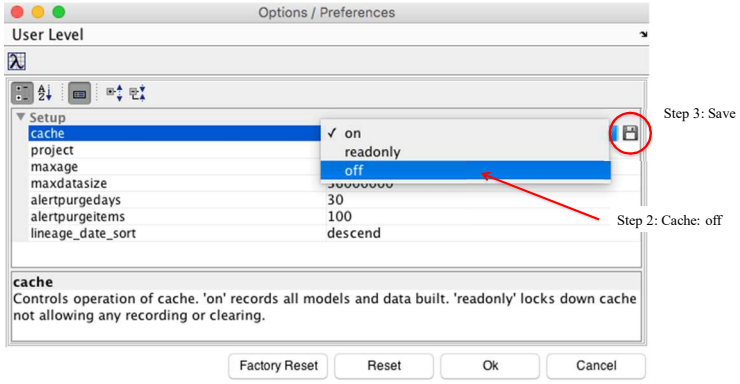
EIGENVECTOR RESEARCH INCORPORATED





Step 1: Edit: Options: Model Cache Settings

EIGENVECTOR RESEARCH INCORPORATED

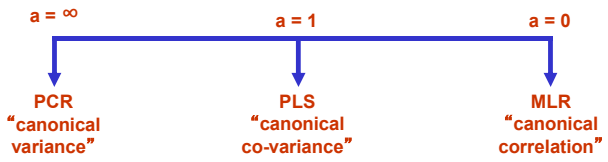


Step 2: Cache: off
Step 3: Save

EIGENVECTOR RESEARCH INCORPORATED

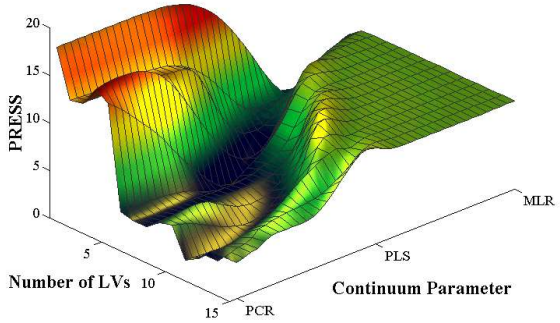
Continuum Regression

- PCR, PLS and MLR can be unified under the single technique Continuum Regression (CR)
- CR is continuously adjustable and encompasses PLS and includes PCR and MLR at the extremes



EIGENVECTOR RESEARCH INCORPORATED

CR Press Surface



EIGENVECTOR RESEARCH INCORPORATED

Missing Data

- MDCHECK
  - Checks data sets for `'NaN'` and `'inf'` and replaces with values consistent with a PCA model (if desired)
    - *e.g.*, see the ISFINITE function
  - This is an iterative method
  - Example, use some data from SFCM

```
>> x = xblock1.data(1:50,[6:9 16:19]);
>> x2 = x(:,2);
>> x(2:4:50,2) = NaN;
```

every 4th sample of column 2 removed

EIGENVECTOR RESEARCH INCORPORATED

Call *MDCHECK*

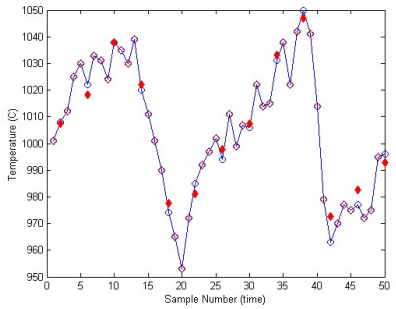
- Change the options to reduce the number of PCs

```
>> options = mdcheck('options')
options =
        max_pcs: 5
       frac_ssq: 0.9500
     meancenter: 'yes'
         output: 'no'
      tolerance: [1.0000e-006 100]
    max_missing: 0.4000
>> options.max_pcs = 3;
>> [flag,mismap,xfill] = mdcheck(x,options);
```


EIGENVECTOR RESEARCH INCORPORATED

MDCHECK Results

```
>> plot(1:50,x2,'ob-',1:50,xfill(:,2),'rd'), hold on
>> plot(2:4:50,xfill(2:4:50,2),'rd','markerfacecolor',[1 0 0])
```



EIGENVECTOR RESEARCH INCORPORATED

Regression Summary

- Regression models can be divided into CLS (used when pure analyte spectra are available) and ILS models (MLR, PCR, PLS, RR, CR, ...)
- PCR and PLS work with ill-conditioned data by reducing to a smaller number of factors
  - has advantage of signal averaging
- Cross-validation is used to determine number of factors
- Fit and Prediction are two different things

EIGENVECTOR RESEARCH INCORPORATED

Model Development

- Developing PCR or PLS models
  - center and scale the data (as appropriate)
  - cross-validate to determine number of factors
  - check **X**-block Q, T^2, leverage, and **Y**-block residuals for outliers
    - remove / explain outliers
  - check RMSEC and RMSECV values for overfit
  - repeat as necessary
- PCR or PLS models consist of
  - mean and scaling vectors
  - **X**-block loadings **P**, scores **T**, and weights **W** (if PLS)
  - **Y**-block loadings **Q**, and scores **U**
  - inner coefficients **b**
  - all of this can be reduced to y = **xb**+a form for prediction with new data

EIGENVECTOR RESEARCH INCORPORATED

Model Application

- A PCR or PLS model is applied by
  - centering and scaling to the model mean and variance
  - multiply measurements by regression vector to get scaled predictions
  - rescale the predictions back to original units using model mean and variance
- Prediction outliers can be found by
  - calculating Q and T^2 values for new samples
- All the modeling and application is packaged:
  - the model is an object that contains all the parameters
  - validation *e.g.*:

```
valid = pcr(x,y,model,options); %pred's with new X- & Y-block
```

  - prediction *e.g.*:

```
pred = pcr(x,model,options);  %predictions with a new X-block
```


EIGENVECTOR RESEARCH INCORPORATED

Model Application – Object Form

- All the modeling and application is packaged:
  - the model is an object that contains all the parameters
  - validation *e.g.*:

```
valid = model.apply(x,y); % this creates a prediction object
pred = model.apply(x);  % this does too

valid_scores = valid.plotscores(options);
pred_scores = pred.plotscores(options);
```

This will create dataset objects containing all of the information that you'll see in a scores plot when using the Analysis interface