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*

```
RECALL
% load plsdata
% x  = delsamps(xblock1,[73 167 278 279]);
% y  = delsamps(yblock1,[73 167 278 279]);

>> pls    %starts a gui: >> regression
```


EIGENVECTOR RESEARCH INCORPORATED

Example PLS using `PLS`

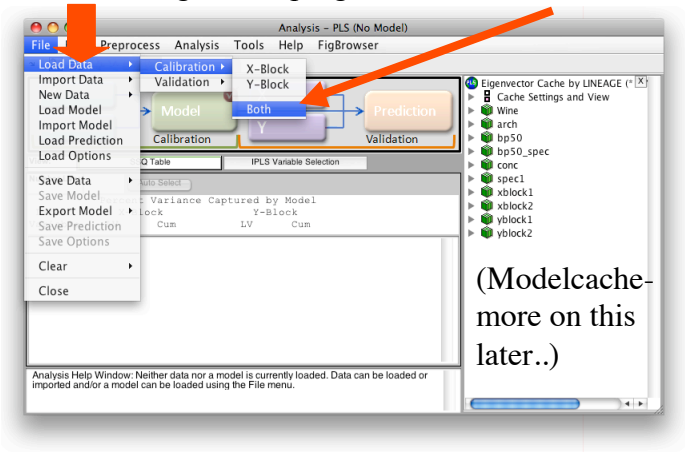
- Construct a linear regression for Y-block1 from X-block1 (time series data)
  - predict level of slurry fed ceramic melter (Y-block)
  - using melter temperatures (X-block)
- Test the model on X-block2 and Y-block2

EIGENVECTOR RESEARCH INCORPORATED

Analysis of plsdata Data

1 Type `pls` at the command prompt » to start the regression program.

2 Click File:Load Data: Calibration:Both menu



EIGENVECTOR RESEARCH INCORPORATED

Load plsdata.mat: xblock1

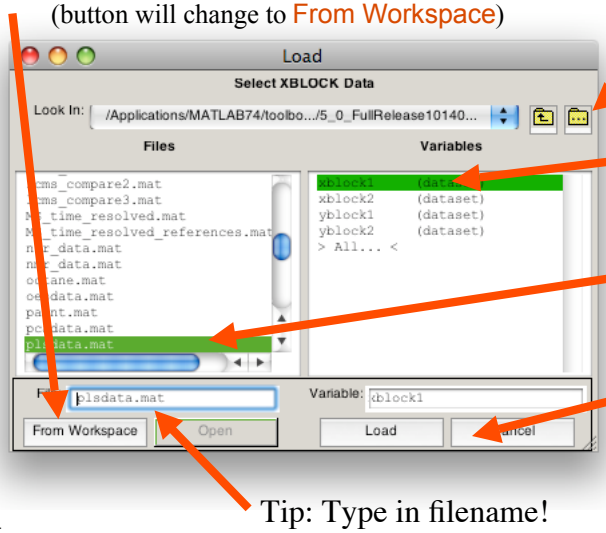
1 Click From File button to load from disk (button will change to From Workspace)

2 Browse to desired folder

4 Highlight xblock1

3 Highlight plsdata.mat

5 Click load

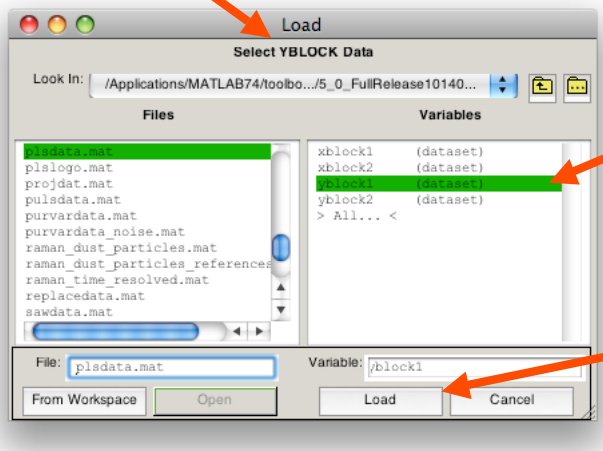


Tip: Type in filename!

EIGENVECTOR RESEARCH INCORPORATED

Load plsdata.mat: yblock1

1 Select Y-Block Data appears automatically



2 Highlight yblock1

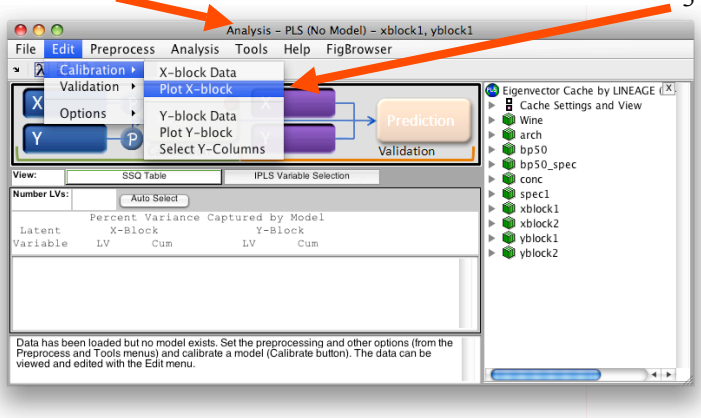
3 Click load

Data: loaded but not analyzed

1 Repeat previous procedure to load validation data (xblock2, yblock2)

2 status window after load

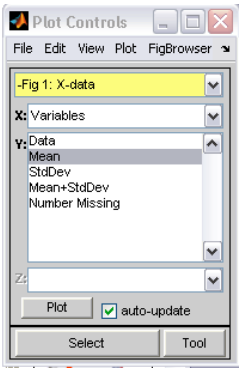
3 Plot your Data: Select Edit:Calibration: Plot X-block

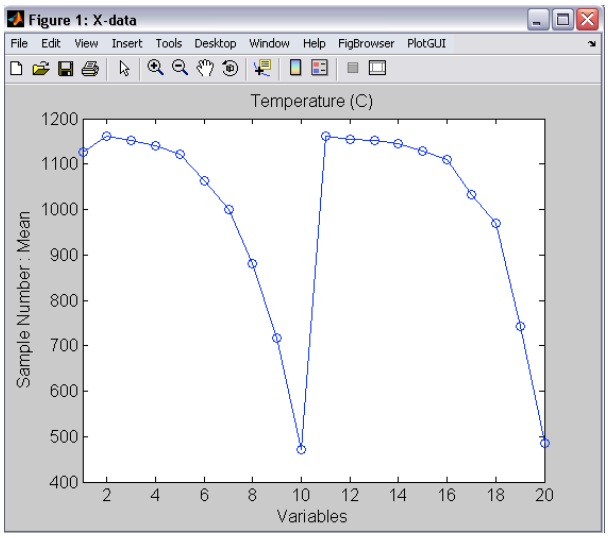


Plot Your Data

Default plot

Column / Variable mean





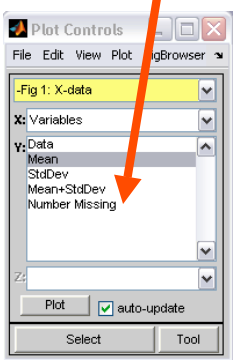
EIGENVECTOR RESEARCH INCORPORATED

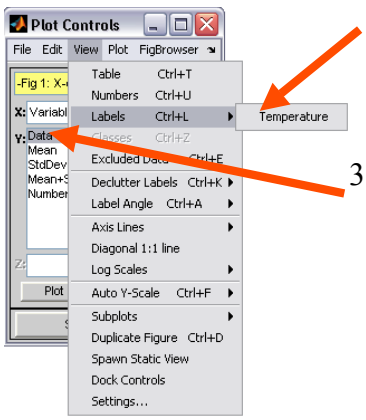
Plot Your Data

1 Plot control default

can look at summary stats

the Plot control generates plots in MATLAB figure windows





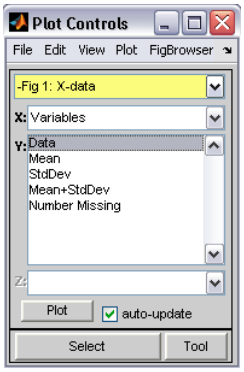
2 under view menu check labels:Temperature

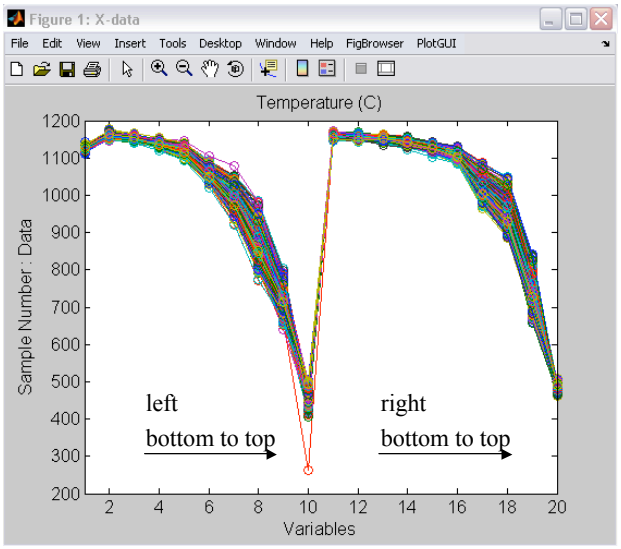
3 under Y: menu highlight Data

EIGENVECTOR RESEARCH INCORPORATED

Plot Your Data

all samples vs. variables



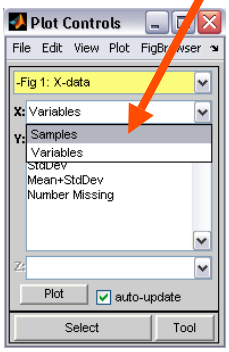


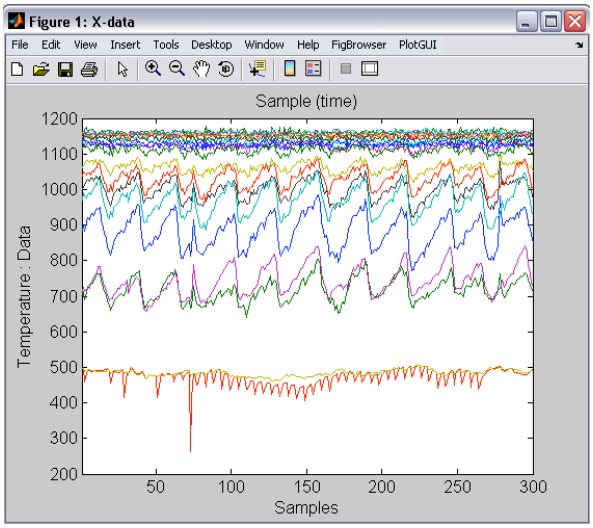
EIGENVECTOR RESEARCH INCORPORATED

Plot Your Data

all variables vs. samples (time)

under X: menu highlight Samples





Also: Edit:Plot Y-Block

EIGENVECTOR RESEARCH INCORPORATED

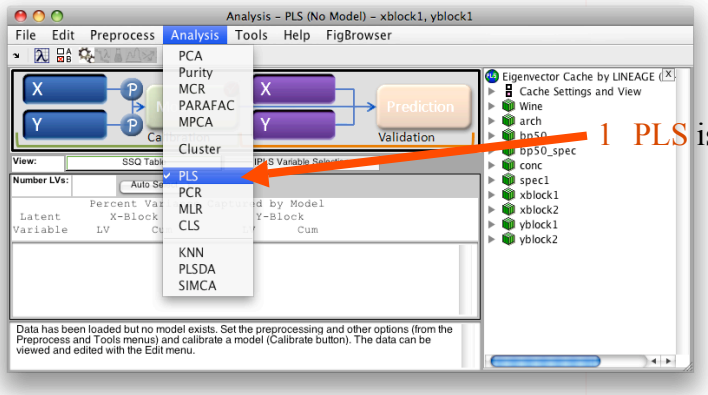
Plot Your Data Summary

- 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
- We'll choose PLS...

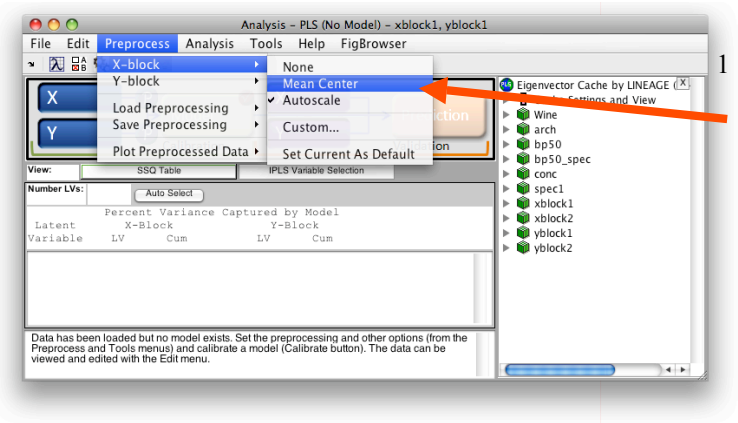


1 PLS is the default

EIGENVECTOR RESEARCH INCORPORATED

How Should We Scale the Data?

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



1 autoscaling is the default, highlight Mean Center

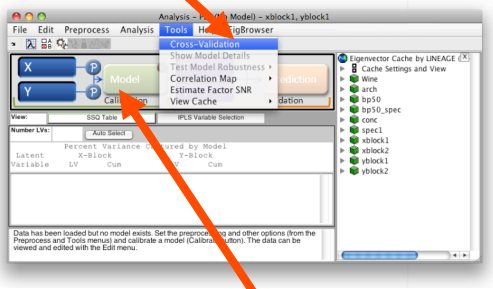
EIGENVECTOR RESEARCH INCORPORATED

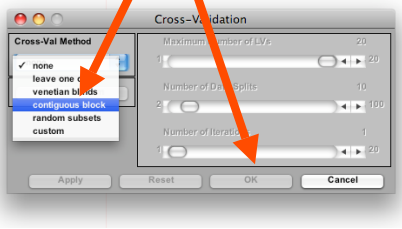
How Should We Cross-Validate?

- Time series data suggests contiguous block cross-validation

1 under Tools menu highlight Cross-Validation

2 contiguous block, then OK





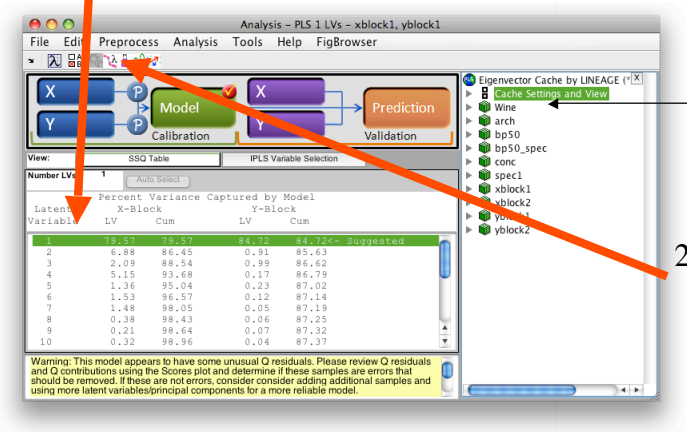
3 click the calculate (gears) or model button to perform the cross-validation and build the regression model

EIGENVECTOR RESEARCH INCORPORATED

Regression Results

1 After calculating the model:

- variance captured table: eigenvalues and % variance explained for each LV.



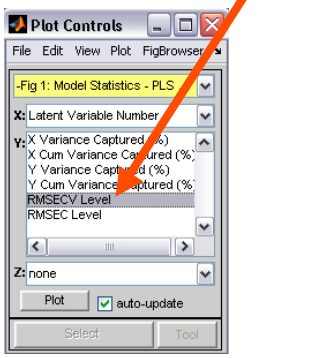
(Check out the modelcache..)

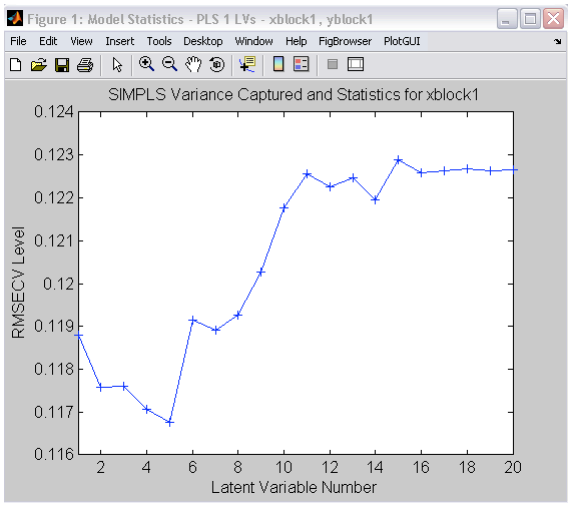
2 Click Eigenvalue button to plot the RMSECV

EIGENVECTOR RESEARCH INCORPORATED

RMSECV Plot

Plot the RMSECV vs. LV.

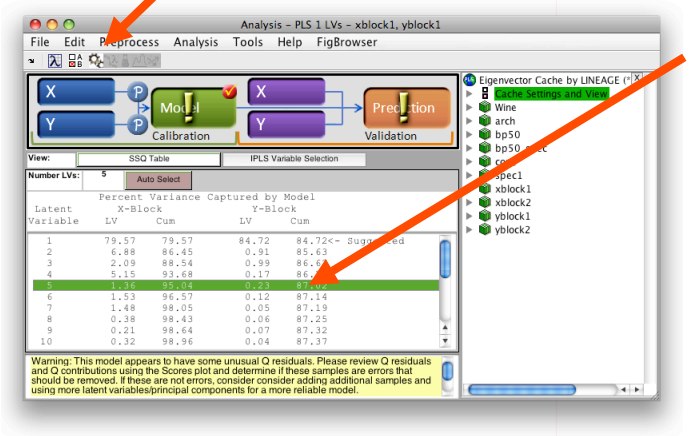




EIGENVECTOR RESEARCH INCORPORATED

Choose Number of LVs

2 Click the Calculate or Model button to reconstruct the model with 5 LVs

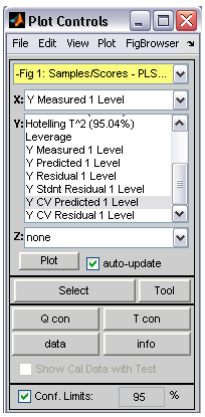


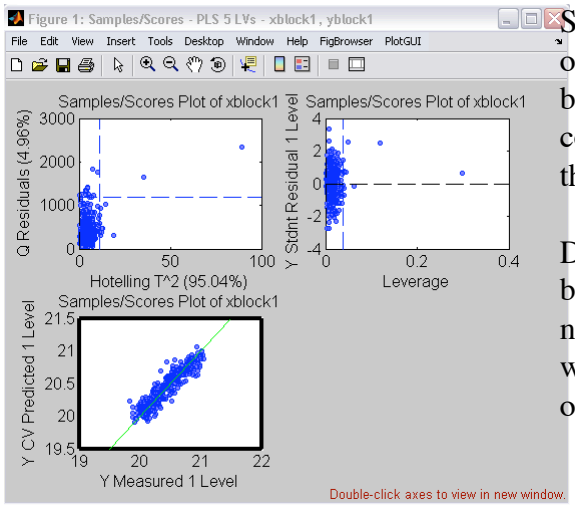
1 Highlight the fifth line to select 5 LVs

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

EIGENVECTOR RESEARCH INCORPORATED

Scores Summary Plot





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

Double-click brings up a new window with that plot only

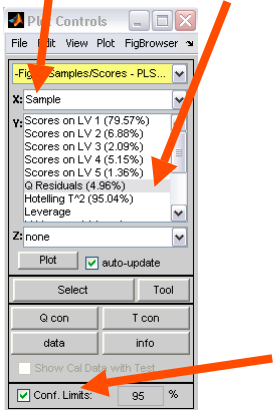
EIGENVECTOR RESEARCH INCORPORATED

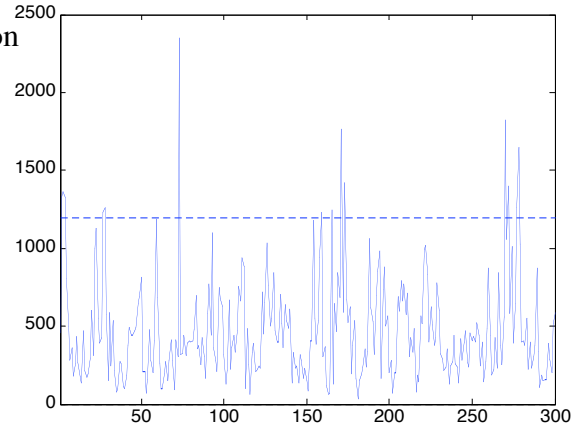
X-Block Q Residuals (by sample)

1 Click Scores Button

X:Sample Scale

Y:Q Residuals





2 Check Conf. Limits

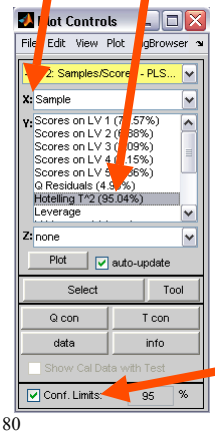
EIGENVECTOR RESEARCH INCORPORATED

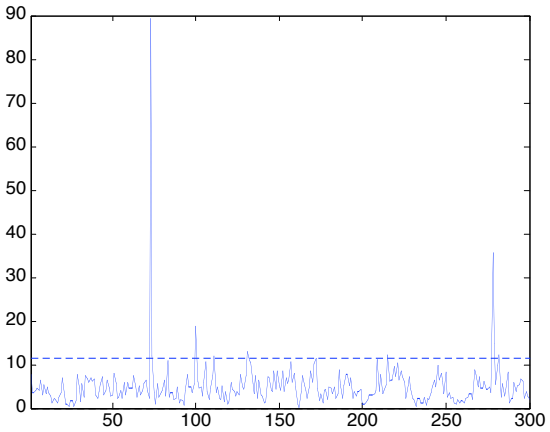
X-Block Hotelling T^2

1 Click Scores Button

X:Sample Scale

Y: Hotelling T^2





2 Check Conf. Limits

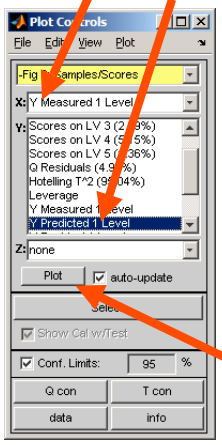
EIGENVECTOR RESEARCH INCORPORATED

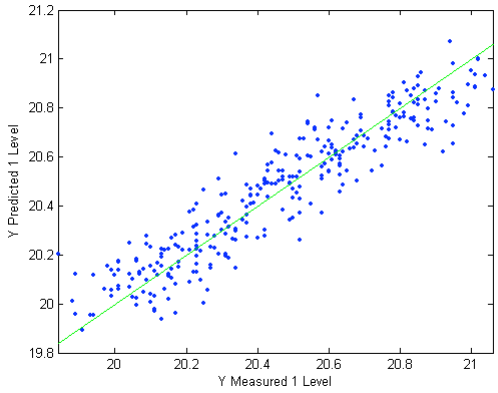
Calibration Curve (Predicted vs. Measured)

1 Click Scores Button

X:Y Measured

Y:Y Predicted





2 Click Plot if auto-update is not checked

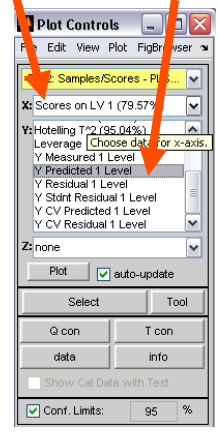
EIGENVECTOR RESEARCH INCORPORATED

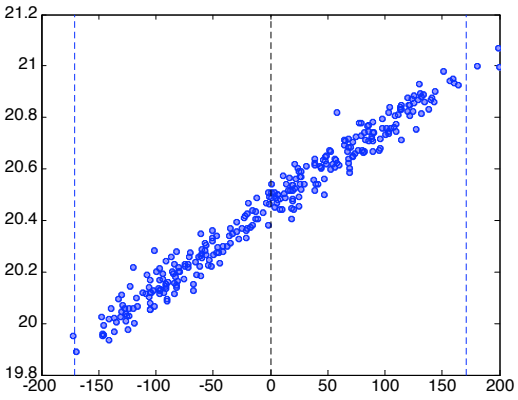
Predicted Y vs. LV 1

1 Click Scores Button

X:LV 1

Y:Y Predicted





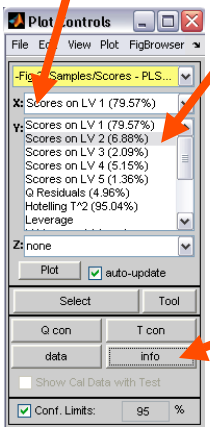
EIGENVECTOR RESEARCH INCORPORATED

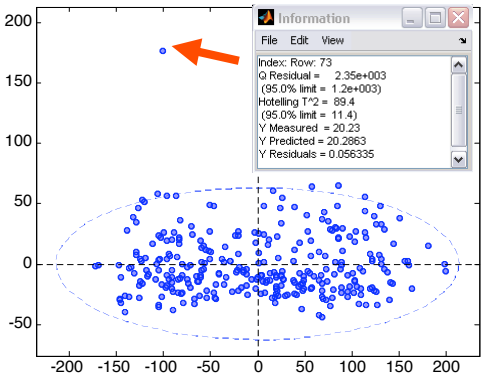
LV 2 vs. LV 1

1 Click Scores Button

X:Scores on LV 1

Y:Scores on LV 2





2 Click info Button and select sample

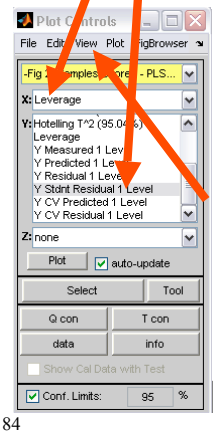
EIGENVECTOR RESEARCH INCORPORATED

Studentized Y Residual vs. Leverage

1 Click Scores Button

X:Leverage

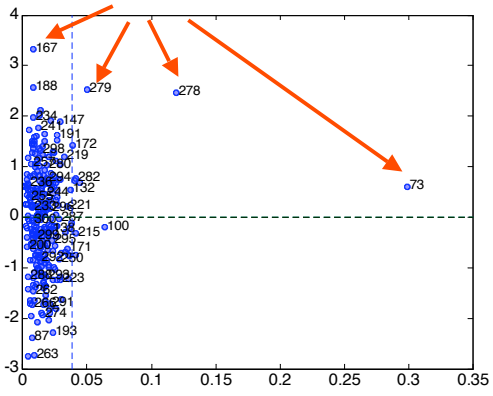
Y:Y Stdnt Residual



2 View: Numbers

3 View: Declutter Labels: Moderate

removed in first example



EIGENVECTOR RESEARCH INCORPORATED

Calculation of Studentized Residuals

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

$$l_i = \mathbf{x}_i \mathbf{X}^+(:,i)$$

- Studentized residuals for column j^{th} of $\mathbf{Y}$, te_j

$$\mathbf{e}_j = \hat{\mathbf{y}}_j - \mathbf{y}_j$$

$$s_j = \left(\frac{1}{m-1} \mathbf{e}_j^T \left((1-\mathbf{l})^T (1-\mathbf{l}) \right)^{-1} \mathbf{e}_j \right)^{1/2}$$

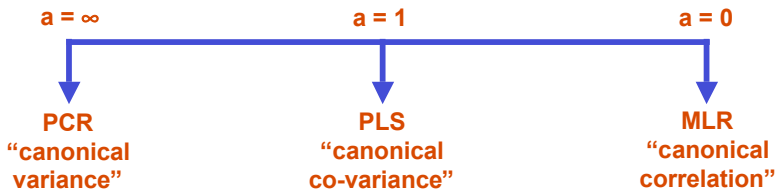
$$\mathbf{t}_{ej} = \mathbf{e}_j ./ \left(s_j (1-\mathbf{l})^{1/2} \right)$$

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 $2nx/m$, where nx is the number of X-block variables

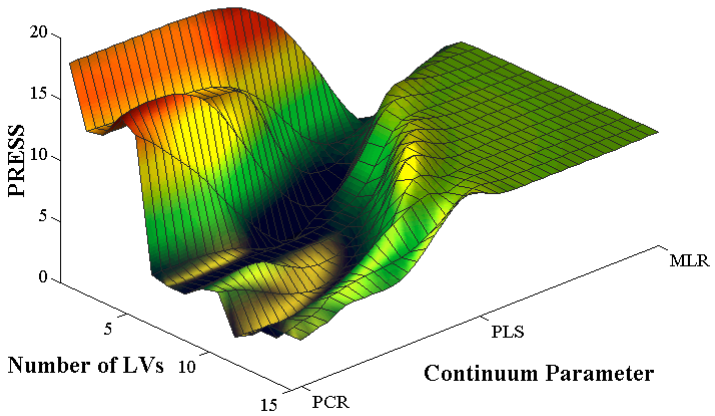
Continuum Regression

- 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

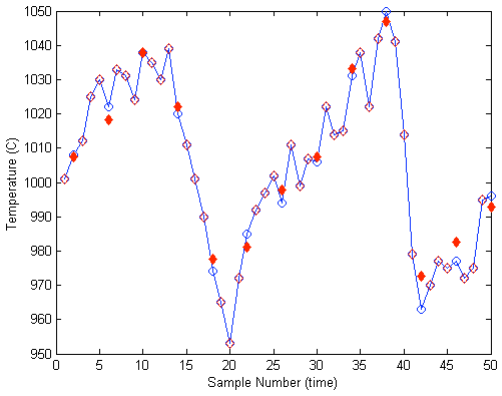
Call MDCHECK

- 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);
```

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 reduces to y = **xb**+a form for prediction with new data

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 a structure 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
```