Alternative Model Forms for Multi-set, Multi-level and Multi-block Data

EIGENVECTOR
RESEARCH INCORPORATED

Outline

- Definitions
- Multi-level data
  - DOE, crossed and nested designs
- ASCA
  - ANOVA simultaneous component analysis
  - Example
- MLSCA
  - Multi-level simultaneous component analysis.
  - Example
- Multi-block Data
  - Levels of data fusion
  - Examples

Definitions

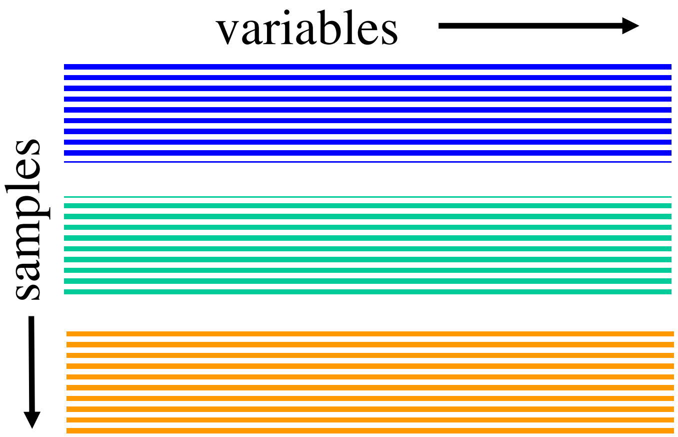
- Single-block: data that is logically contained in a single matrix
- Two-block: two single block data sets that share a common mode (typically the sample mode)
- Multi-block: multiple single blocks that share a common mode
- Multi-set: groups of related samples that have the same variables, typically from designed experiments
- Multi-level: same as multi-set except typically from nested or happenstance designs

Definitions (cont.)

- Multi-way: Data that is logically arranged in 3-way (or more) arrays
- Data fusion: the process of combining multiple sources of data to improve accuracy

Multi-set Data

- Groups (sets) of related samples which have the same variables.



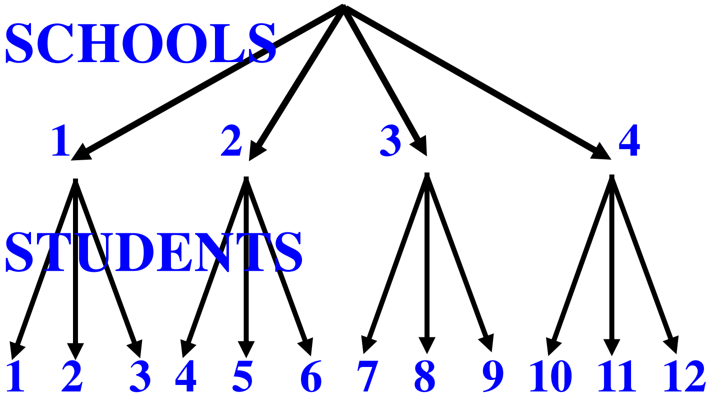
- Differences between groups may hide variability inherent to all samples.
- For samples grouped according to a DoE can separate variability
  - Due to each factor
  - Remaining systematic variability
- This is the purpose of ASCA and MLSCA

Crossed and nested designs

- Crossed (factorial) designs: One or more factors with samples measured for every combination of factor levels.

		Treatment			
		A	B	C	D
Dose	1.1				
	2.0				
	3.5				

- Nested designs: samples belong to groups which are organized hierarchically.



These are both 2-factor designs

ASCA

ANOVA Simultaneous Component Analysis

For multivariate datasets based on crossed experimental designs, ASCA applies ANOVA decomposition and dimension reduction (PCA) to :

- Separate the variability associated with each factor.
- Estimate contribution of each factor to total variance.
- Test main factor and interaction effects for significance.
- View scores and loadings for these effects.

Especially useful for high-dimension datasets where traditional ANOVA is not possible.

ASCA Method

- X data matrix, with 2 factors A and B.
- Decompose into DOE components

$$\mathbf{X} = \mathbf{X}_{\mathrm{avg}} + \mathbf{X}_{\mathrm{A}} + \mathbf{X}_{\mathrm{B}} + \mathbf{X}_{\mathrm{AB}} + \mathbf{E}$$

- Build PCA model for each main effect and interaction

$$\mathbf{X} = \mathbf{X}_{\mathrm{avg}} + \mathbf{T}_{\mathrm{A}}\mathbf{P}_{\mathrm{A}}^{\mathrm{T}} + \mathbf{T}_{\mathrm{B}}\mathbf{P}_{\mathrm{B}}^{\mathrm{T}} + \mathbf{T}_{\mathrm{AB}}\mathbf{P}_{\mathrm{AB}}^{\mathrm{T}}$$

- Calculate permutation P-value to estimate each factor's significance.
- Project residuals onto each PCA sub-model.

ASCA Demo data: asca_data

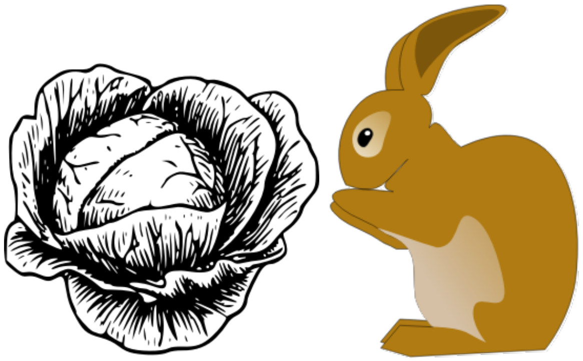
X: Measured glucosinolate levels in cabbage plants,

3 treatments, Control, Root, Shoot.

4 time points, Days 1, 3, 7, and 14.

5 replicates for each time-treatment.

11 measured concentrations.



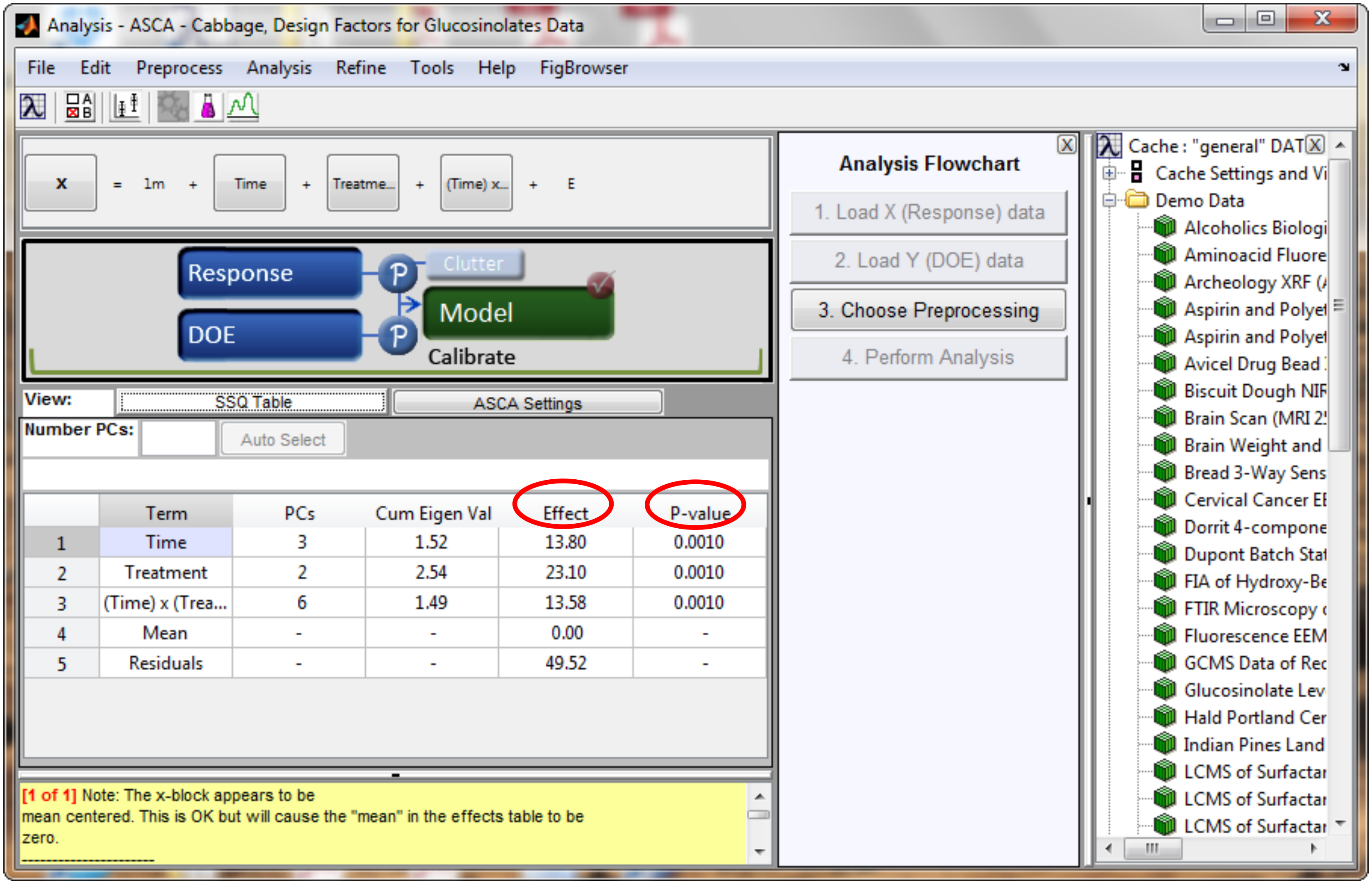
X: (60, 11)

F: (60, 2) design matrix.

See X.description for details.

		Time (Day)			
		1	3	7	14
Treatment	C				
	R	5 replicates each			
	S				

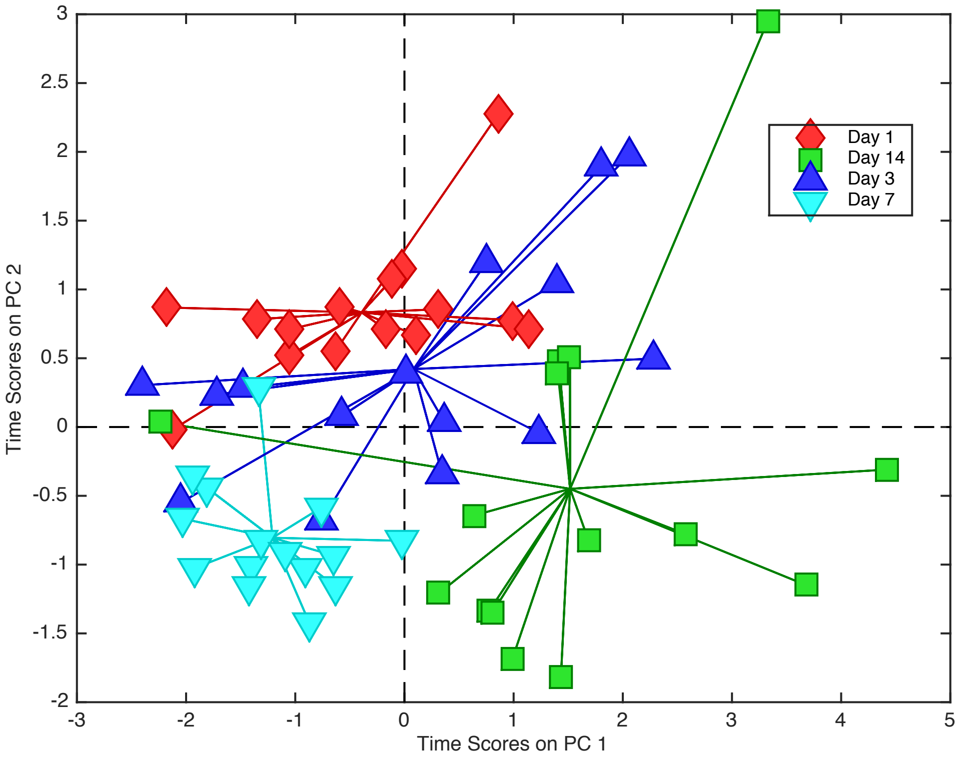
ASCA Model

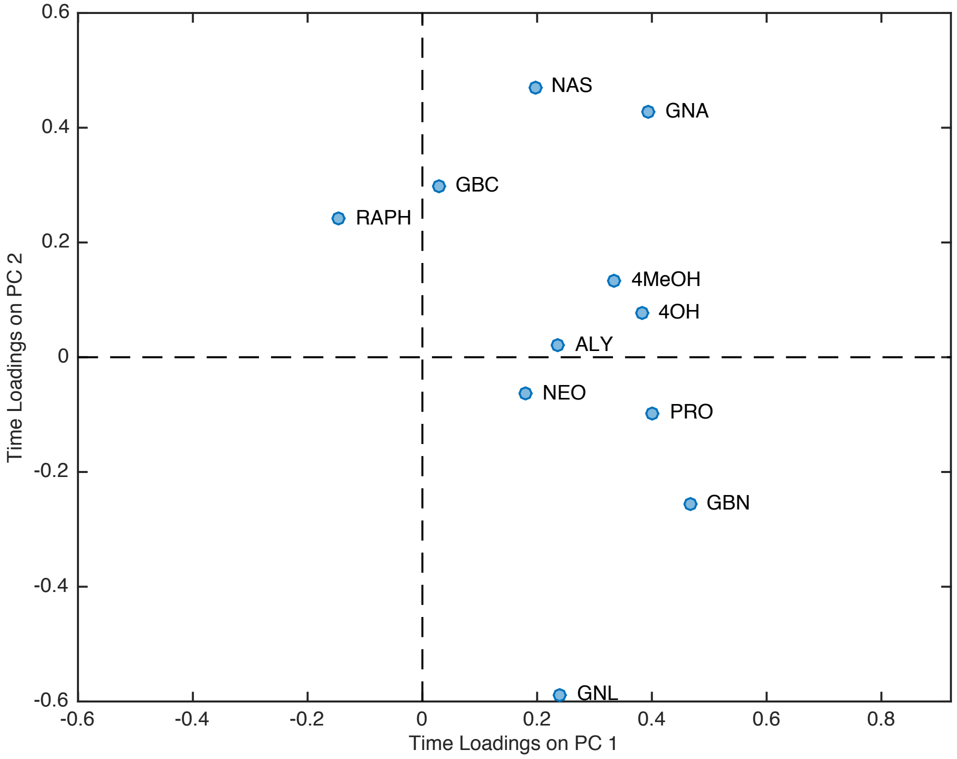

Analysis - ASCA - Cabbage, Design Factors for Glucosinolates Data

	Term	PCs	Cum Eigen Val	Effect	P-value
1	Time	3	1.52	13.80	0.0010
2	Treatment	2	2.54	23.10	0.0010
3	(Time) x (Trea...	6	1.49	13.58	0.0010
4	Mean	-	-	0.00	-
5	Residuals	-	-	49.52	-

[1 of 1] Note: The x-block appears to be
mean centered. This is OK but will cause the "mean" in the effects table to be
zero.

Time Model Scores and Loadings

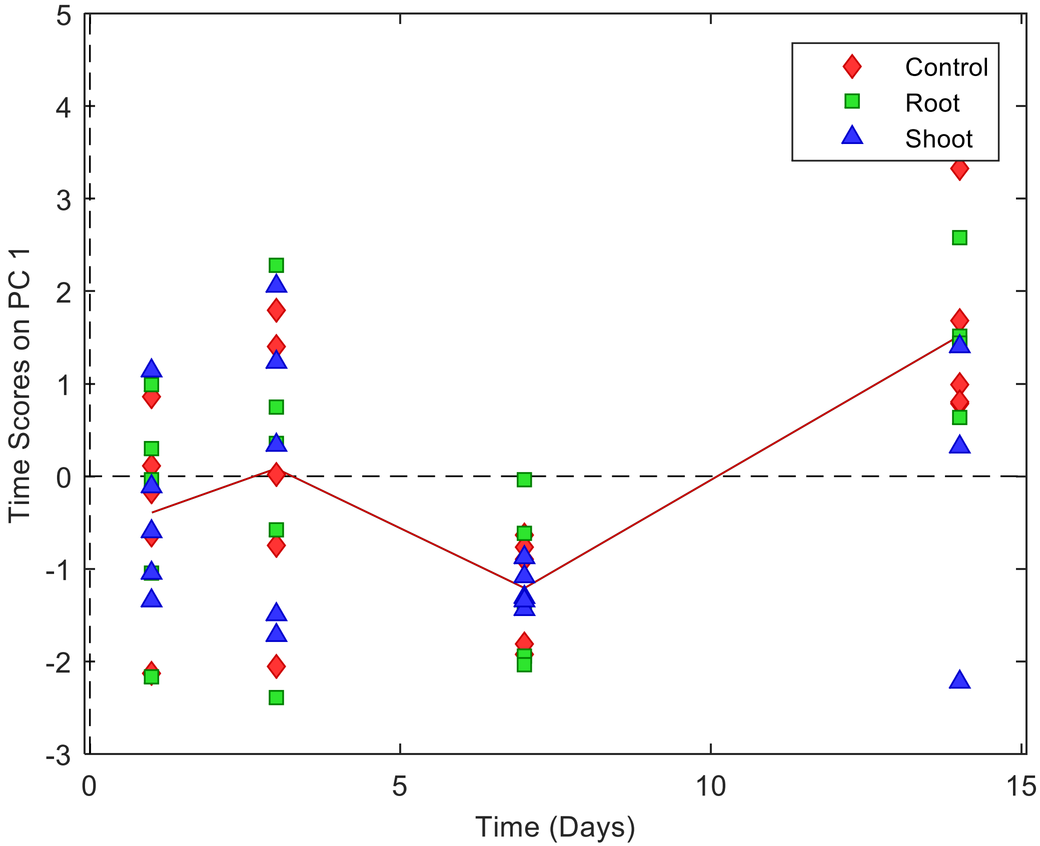




EIGENVECTOR RESEARCH INCORPORATED

ASCA Scores Plot

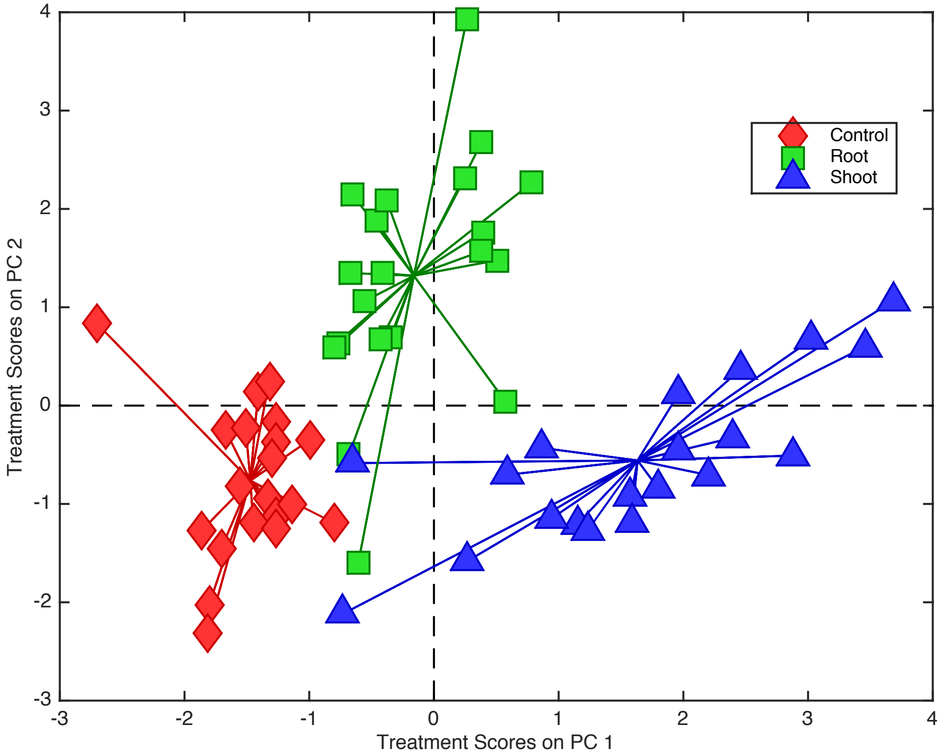
”Time” factor sub-model, PC 1

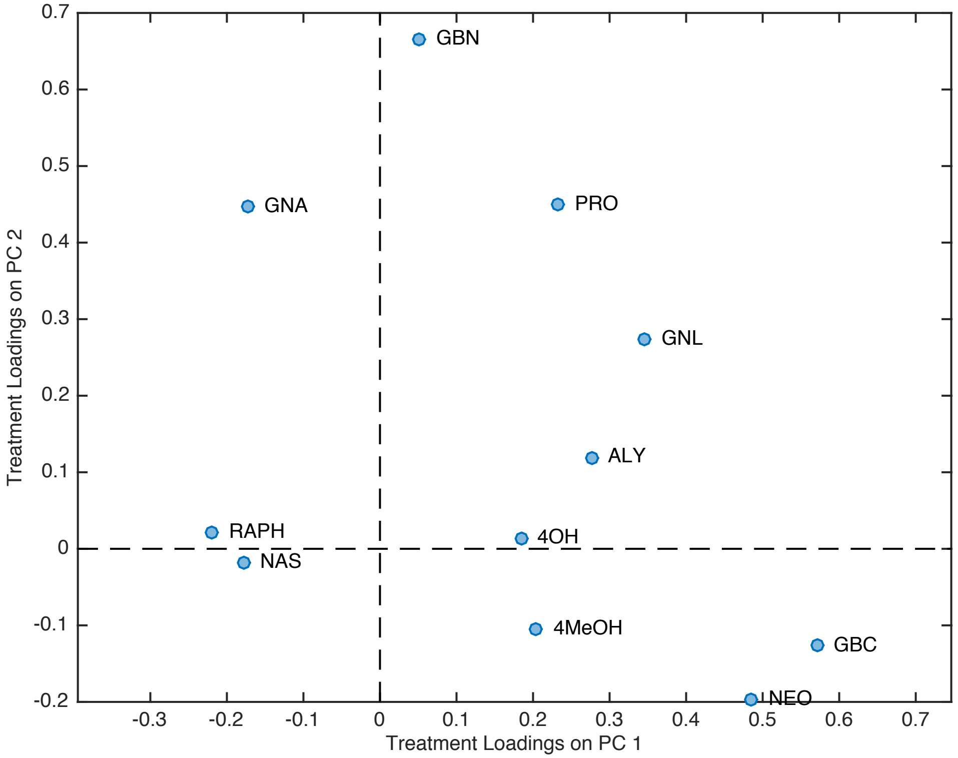


PC 1 of Time dependency common to all Treatments.

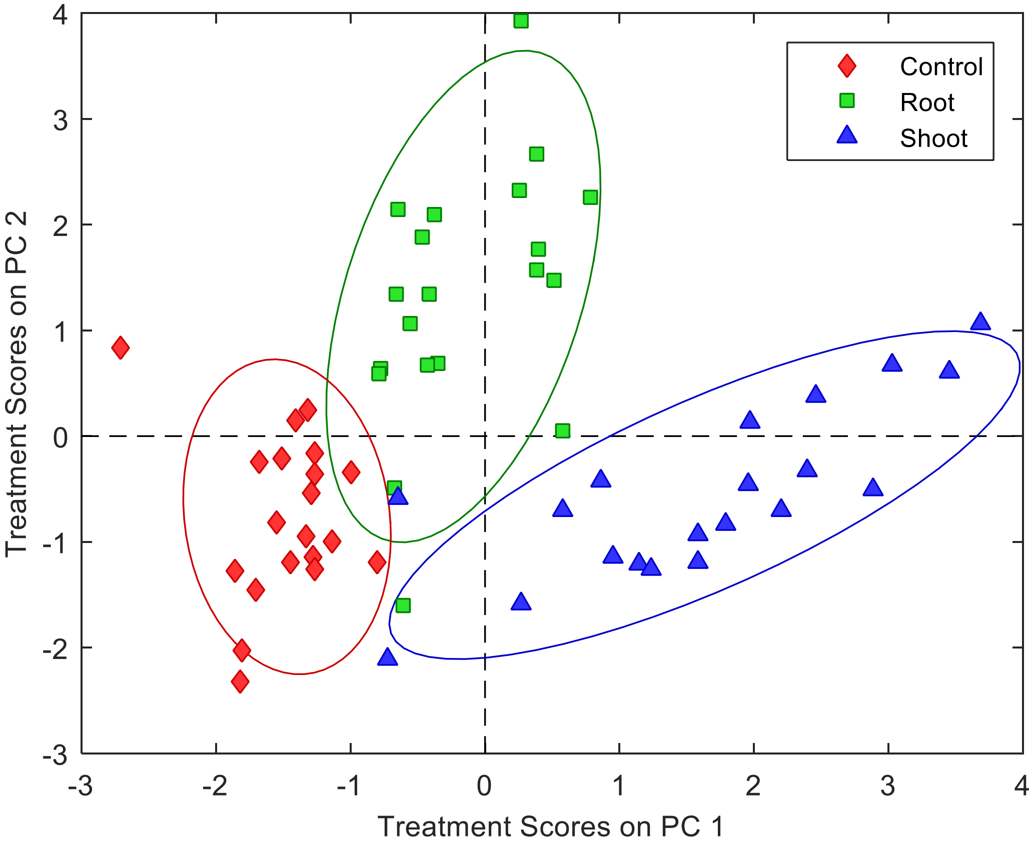
Class = Treatment. Connect Classes = Mean at each X

Treatment Model Scores and Loadings



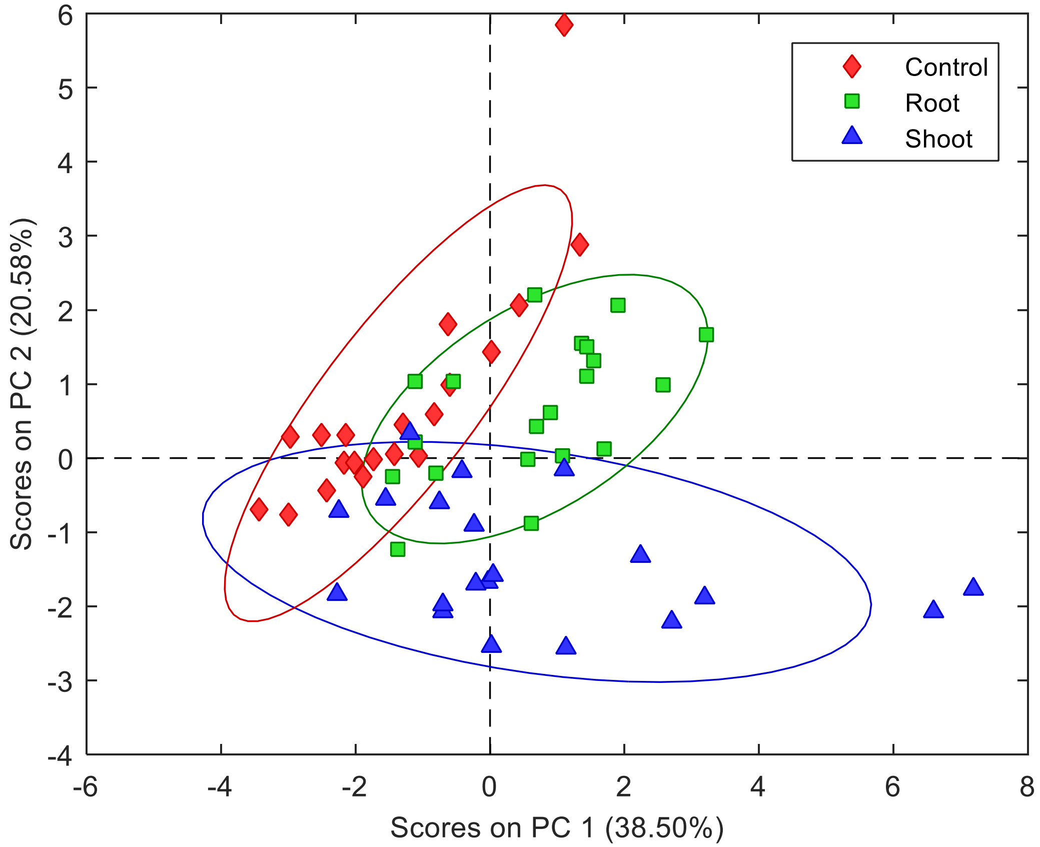


ASCA Treatment Scores Plot



Separating out the Time and Time x Treatment effects highlights the Treatment effect

PCA Scores Plot



…better than is seen by simply applying PCA to the data.

ASCA Conclusions

- ASCA allows the variation associated with each factor to be resolved, and to see the main variables involved.
- For a perturbed biological system
  - Time factor scores reveal the common response independent of Treatment
  - Treatment factor scores show the Treatment effect independent of Time
  - Time x Treatment interaction scores show the additional time dependency at each Treatment level.

ASCA Conclusions, cont.

- The % contribution of each factor or interaction to the total SSQ shows which effects are important.
- Perturbation P-values for each factor estimates the probability that there is no difference between the factor level averages for this effect.

MLSCA

Multi-level Simultaneous Component Analysis

MLSCA is a special case of ASCA applied to data from designed experiments with nested factors.

- Separates variability associated with each factor and residual.
- Estimate contribution of each factor to total sum of squares.
- View scores and loadings for these effects.
- Also builds PCA model on the residuals, or “within” variability. “Within” is often the focus of the analysis.
- Note that “Class Center” pre-processing can achieve same result if there is a single nesting factor.

MLSCA Method

- X data matrix, with 2 nested factors A and B.
- Decompose into DOE components

$$\mathbf{X} = \mathbf{X}_{\mathrm{avg}} + \mathbf{X}_{\mathrm{A}} + \mathbf{X}_{\mathrm{B(A)}} + \mathbf{E}$$

$\mathbf{X}_{\mathrm{A}}$ contains factor A level averages
$\mathbf{X}_{\mathrm{B(A)}}$ contains factor B level averages for each level A
$\mathbf{E}$ are the residuals, “within” component

- Build PCA model for each effect and residual

$$\mathbf{X} = \mathbf{X}_{\mathrm{avg}} + \mathbf{T}_{\mathrm{A}}\mathbf{P}_{\mathrm{A}}^{\mathrm{T}} + \mathbf{T}_{\mathrm{B(A)}}\,\mathbf{P}_{\mathrm{B(A)}}^{\mathrm{T}} + \mathbf{T}_{\mathrm{E}}\mathbf{P}_{\mathrm{E}}^{\mathrm{T}}$$

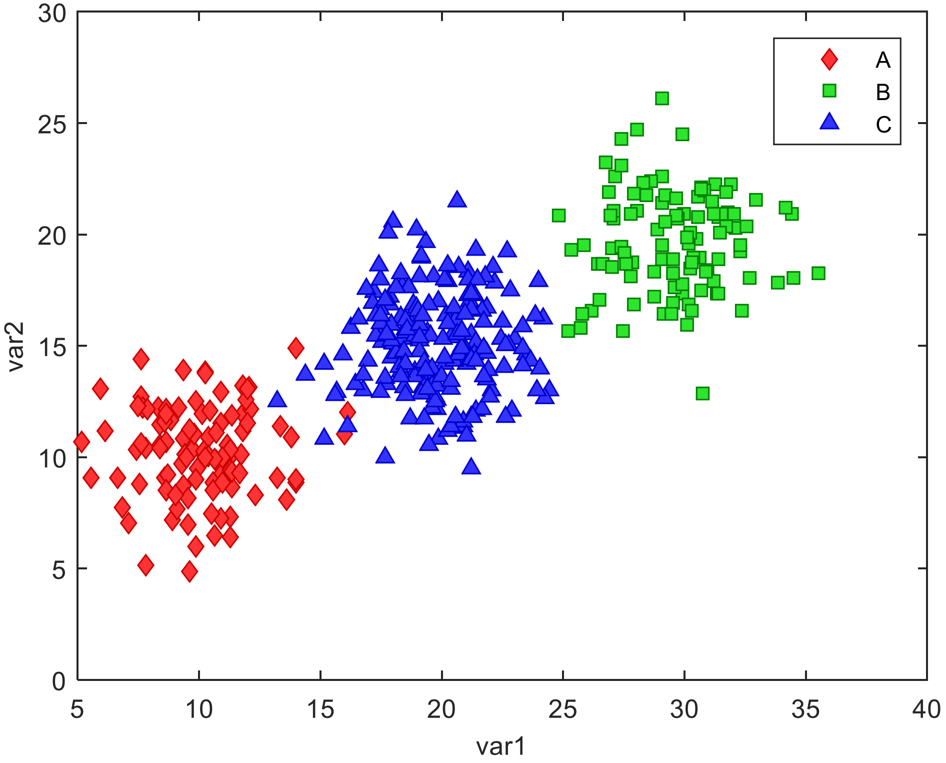
constant *between A* *between B* *within*

MLSCA: simple example

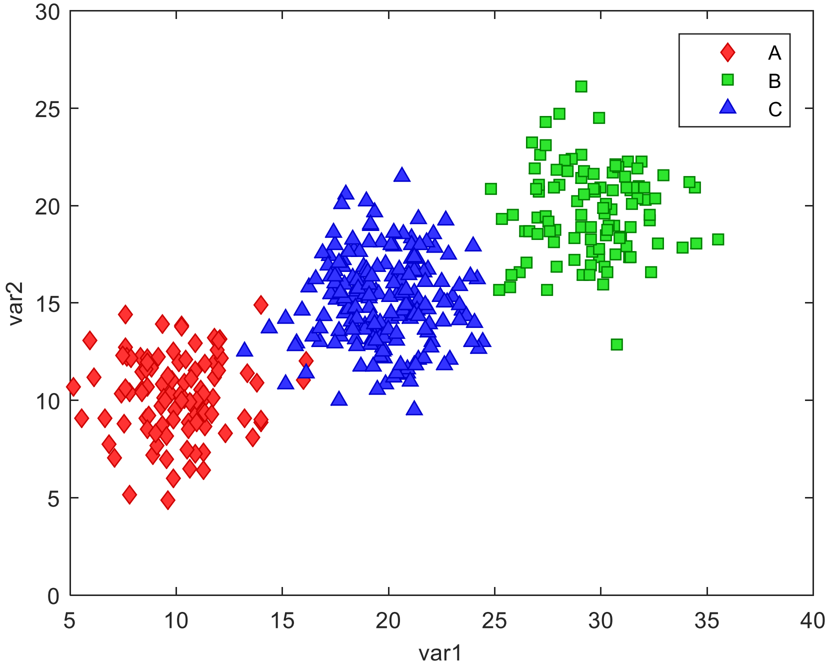
MLSCA can be used to reveal systematic variability within grouped samples which can be obscured by inter-group differences.

Example: X: (400,2)
400 samples from 3 individuals, A, B, and C.

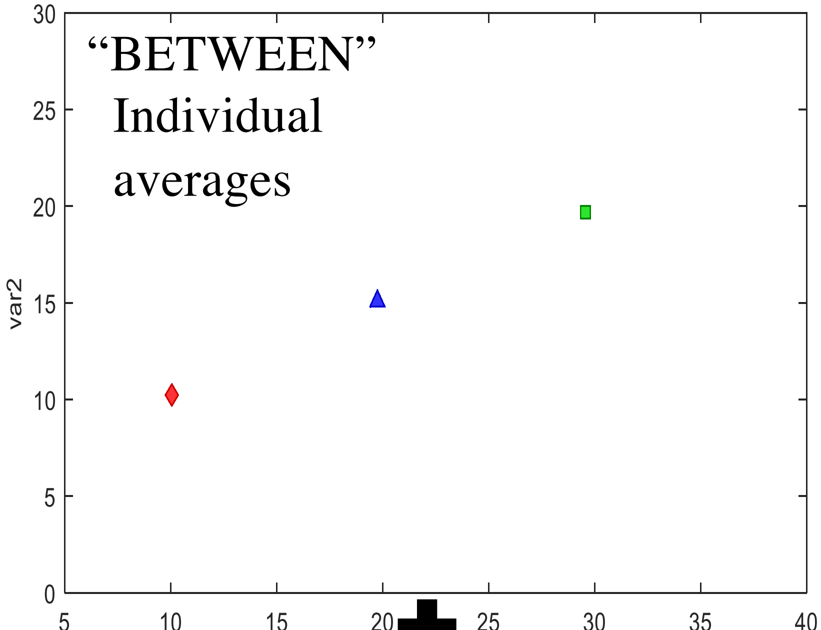
Need to remove offsets for each individual to see the internal, “within” individual variation.



EIGENVECTOR RESEARCH INCORPORATED

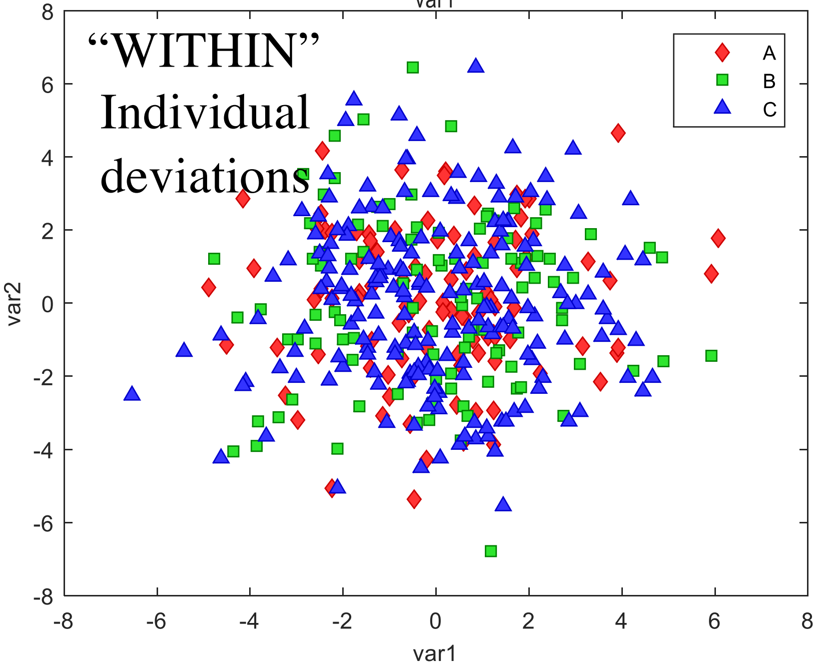


=


“BETWEEN”

Individual averages

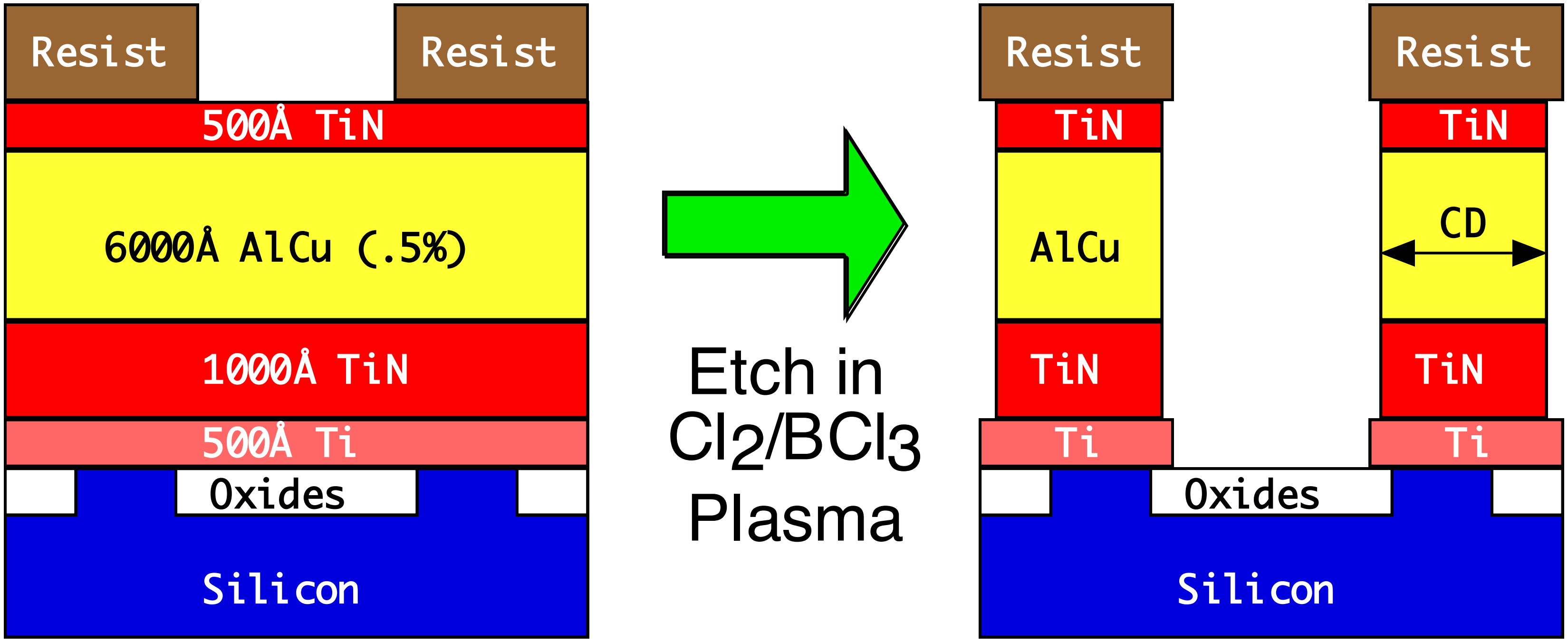
+


“WITHIN”

Individual deviations

X = average for each individual
+ deviations from that

Example: Plasma Metal Etch



- Linewidth (Critical Dimension) Control
  - Constant linewidth reduction run to run and across wafer
  - Constant linewidth reduction for every material in stack
- Minimal damage to oxide

Available Measurements

- Machine State Data: Equipment has SECS-II Port
  - Provides traces with time stamp and step number
  - Regulatory controller setpoints & controlled variable measured values
    - gas flows, pressure, plasma powers
  - Regulatory controller manipulated variables
    - exhaust throttle valve, capacitors
    - mass flow controller do not provide valve position
  - Additional process measurements
    - broadband plasma emission (often used for endpoint)
    - impedance measurements
- Optical Emission Spectroscopy (OES)
- RF Data

Nested dataset "mlsca_data"

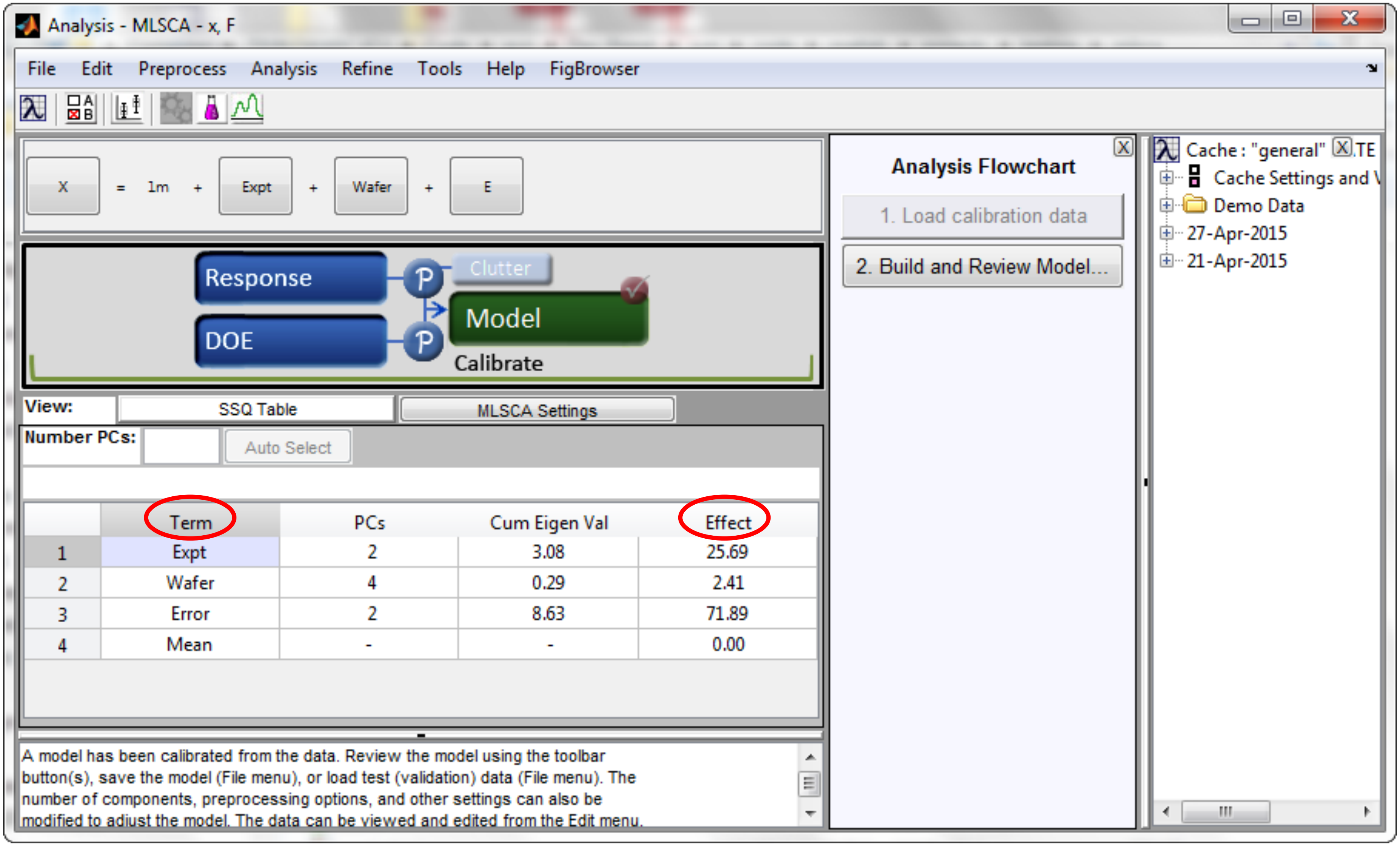
12 engineering variables from a LAM 9600 Metal Etcher over the course of etching 107 wafers.

- Three experiments were run at different times.
- Experiment have 34, 36 and 37 wafers each, for 107 unique wafers.
- 80 samples (replicates) measured for each wafer during etching.
- X is (8560, 12)

	EXPERIMENT											
	1				2				3			
WAFER	1	2	•••	34	35	36	•••	70	71	72	•••	107
80 REPLI-CATES	X X X . . . X	X X X . . . X		X X X . . . X	X X X . . . X	X X X . . . X		X X X . . . X	X X X . . . X	X X X . . . X		X X X . . . X

Nested factors are not crossed.

MLSCA Model



View: SSQ Table

Number PCs:

	Term	PCs	Cum Eigen Val	Effect
1	Expt	2	3.08	25.69
2	Wafer	4	0.29	2.41
3	Error	2	8.63	71.89
4	Mean	-	-	0.00

A model has been calibrated from the data. Review the model using the toolbar button(s), save the model (File menu), or load test (validation) data (File menu). The number of components, preprocessing options, and other settings can also be modified to adjust the model. The data can be viewed and edited from the Edit menu.

MLSCA Scores Plot

”Experiment” factor sub-model, PC 1 vs 2

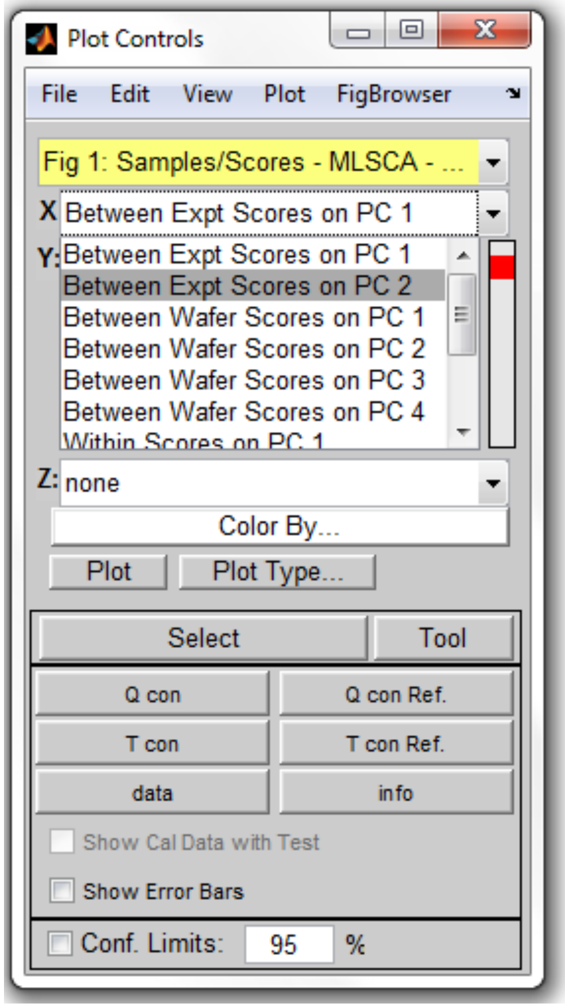

Plot Controls
File
Edit
View
Plot
FigBrowser
Select
Tool
Q con
Q con Ref.
T con
T con Ref.
data
info
□ Show Cal Data with Test
□ Show Error Bars
Color By...
Plot
Plot Type...
Between Expt Scores on PC 1
Between Expt Scores on PC 2
Between Wafer Scores on PC 1
Between Wafer Scores on PC 2
Between Wafer Scores on PC 3
Between Wafer Scores on PC 4
Within Scores on PC 1

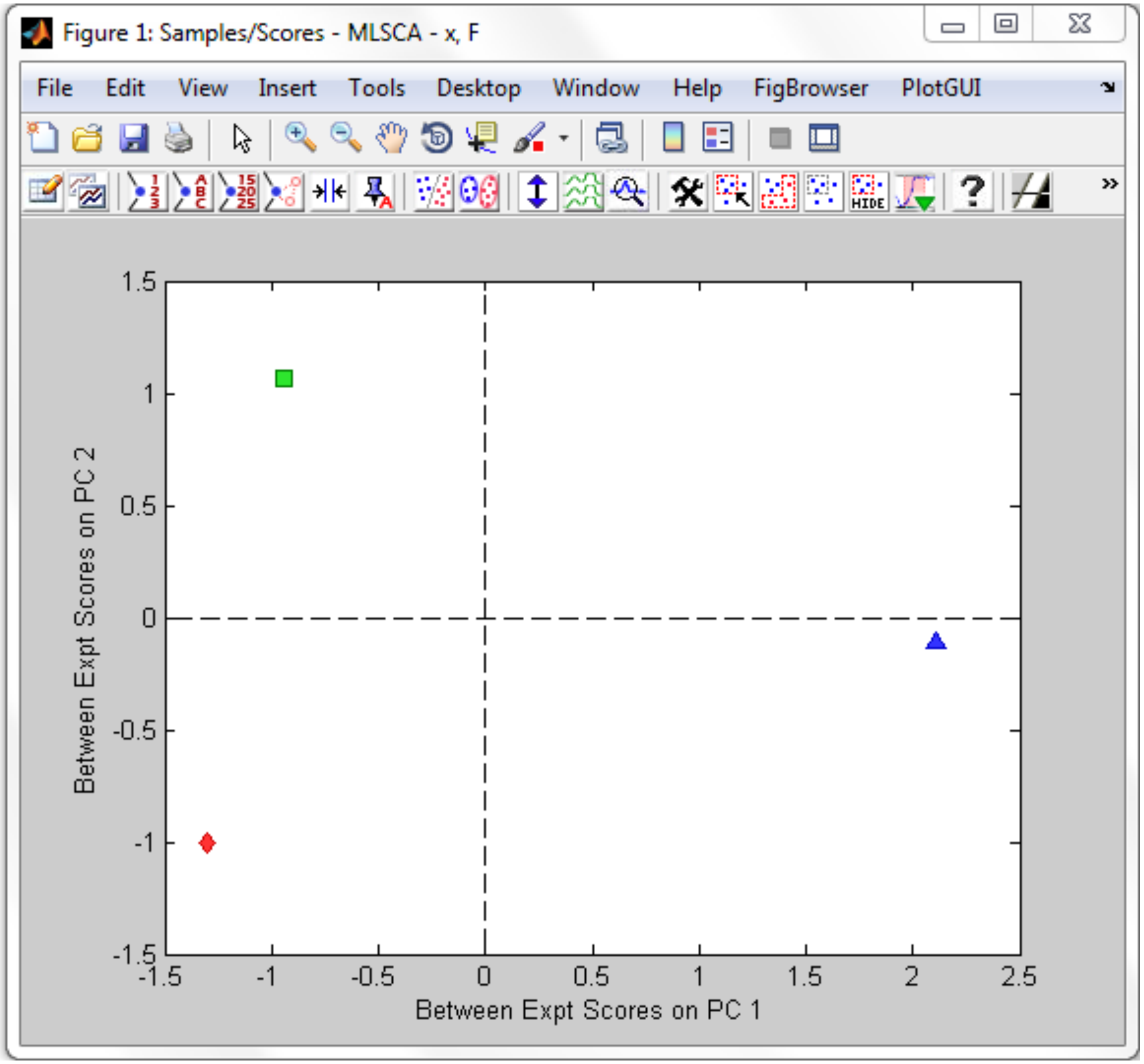
Fig 1: Samples/Scores - MLSCA - ...

X Between Expt Scores on PC 1

Y: Between Expt Scores on PC 2

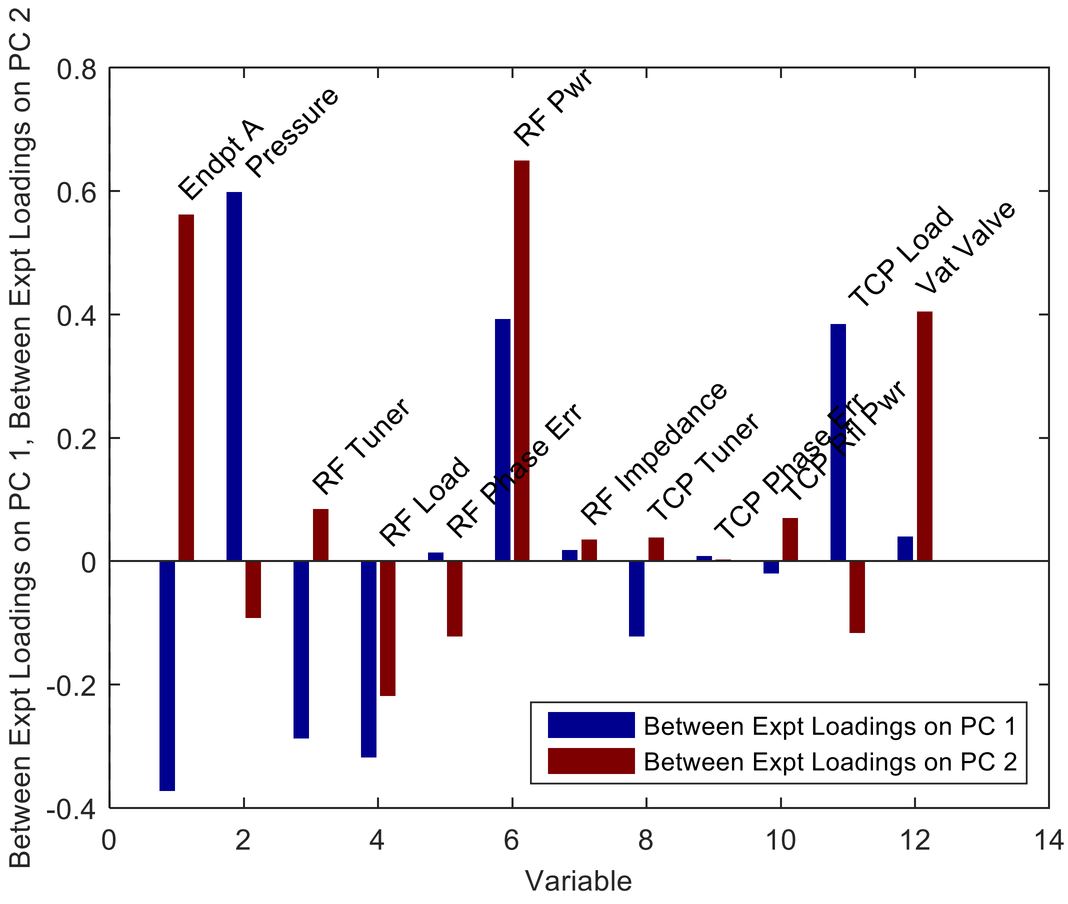
Z: none

□ Conf. Limits: 95 %



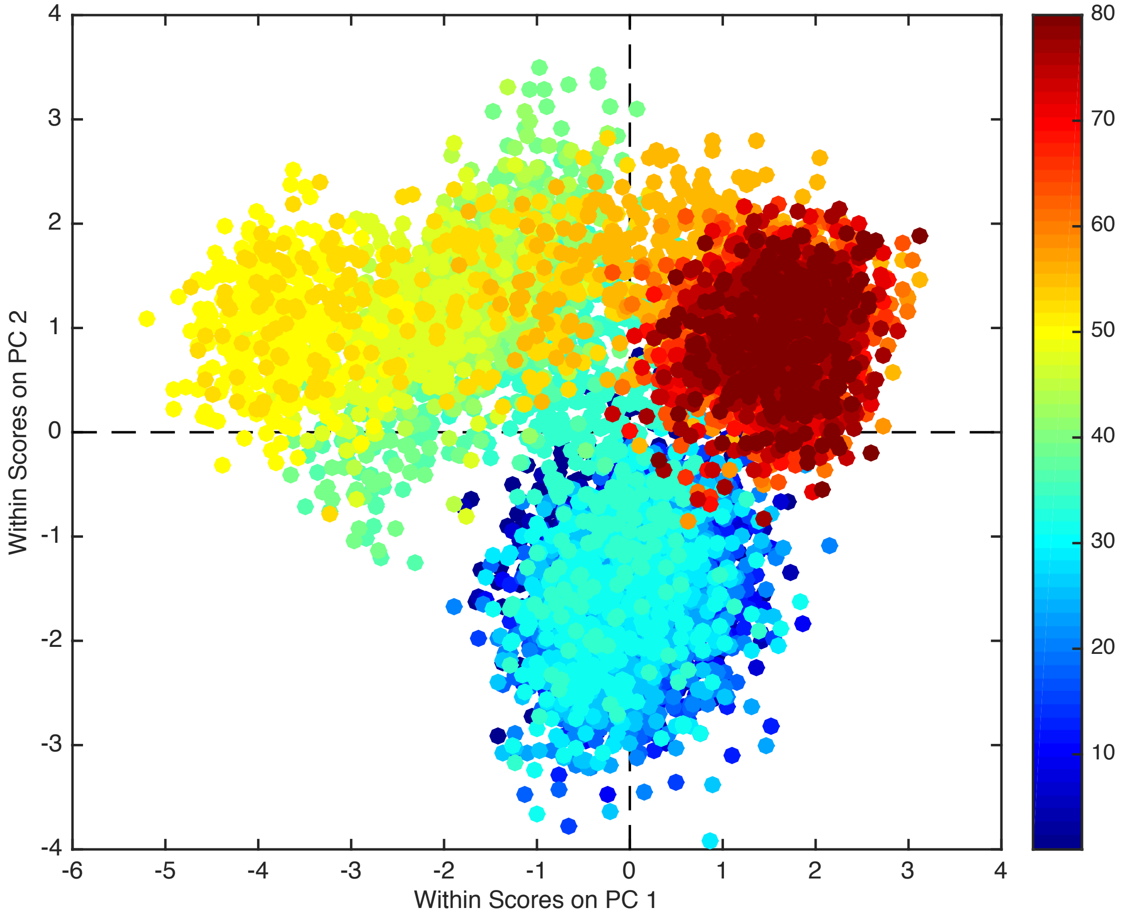
MLSCA Loadings Plot

”Experiment” factor sub-model, PC 1 and 2



MLSCA Scores Plot

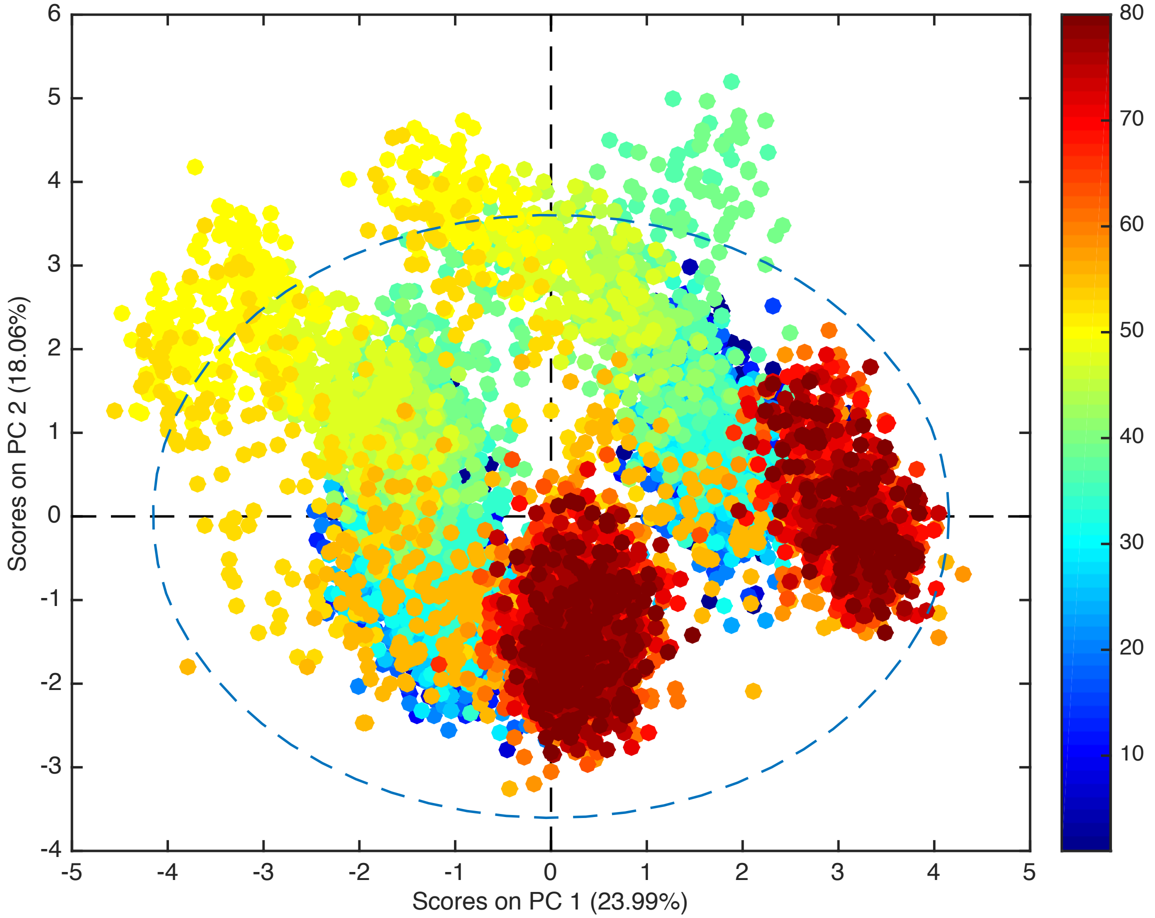
”Within” sub-model, PC 1 vs 2, colored by time



EIGENVECTOR RESEARCH INCORPORATED

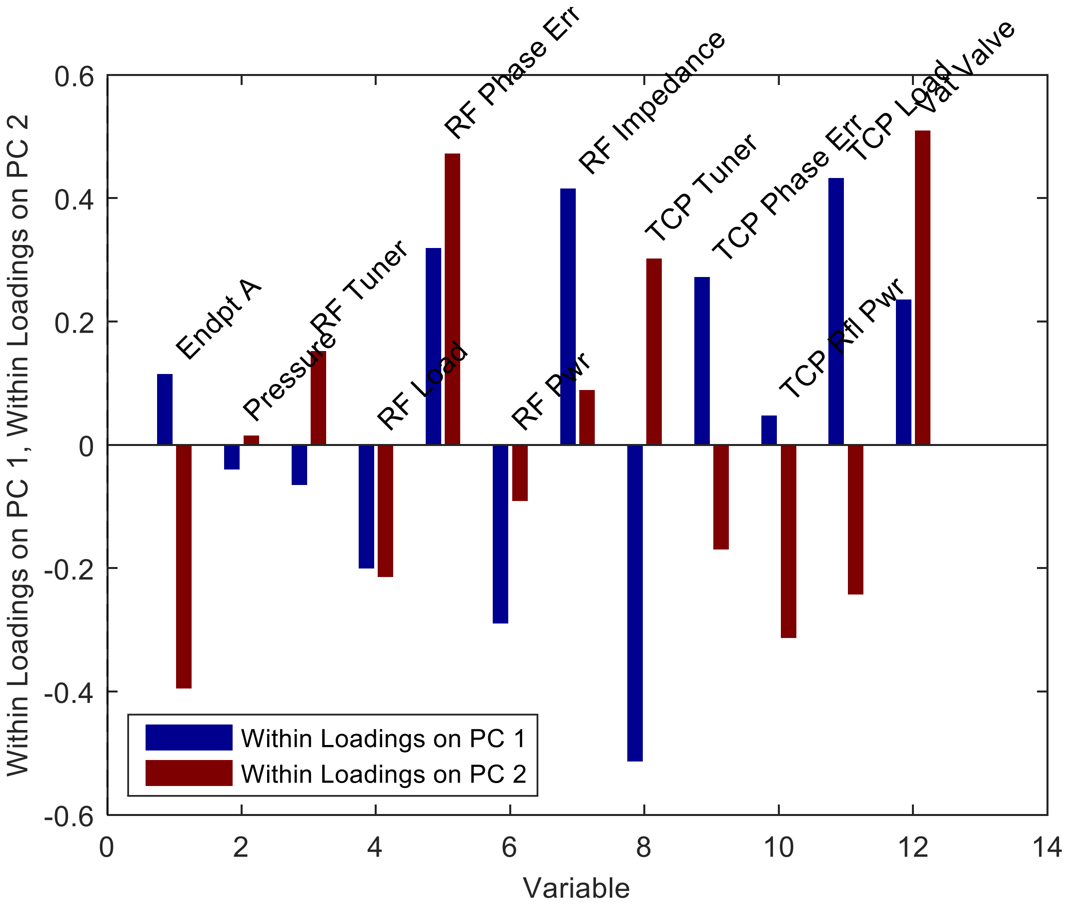
Compare to PCA

Convolves between and within factors



MLSCA Loadings Plot

”Within” Residual sub-model, PC 1 and 2



MLSCA Conclusions

MLSCA allows the variation associated with each nested factor to be resolved, and to see the main variables involved.

- Often used to reveal the inherent “within” group variability of samples after factor effects are removed. For process data this allows separation of within-run variation from between-run variation.
- SSQ contributions show which nested factors are important.

Multi-block Data Fusion

- Data fusion can be done at three levels
  - Low level: single model of combined data blocks appropriately scaled/preprocessed
  - Mid level: combining scores from individual data blocks into a consensus model
  - High level: combining predictions from individual models in some sort of voting scheme

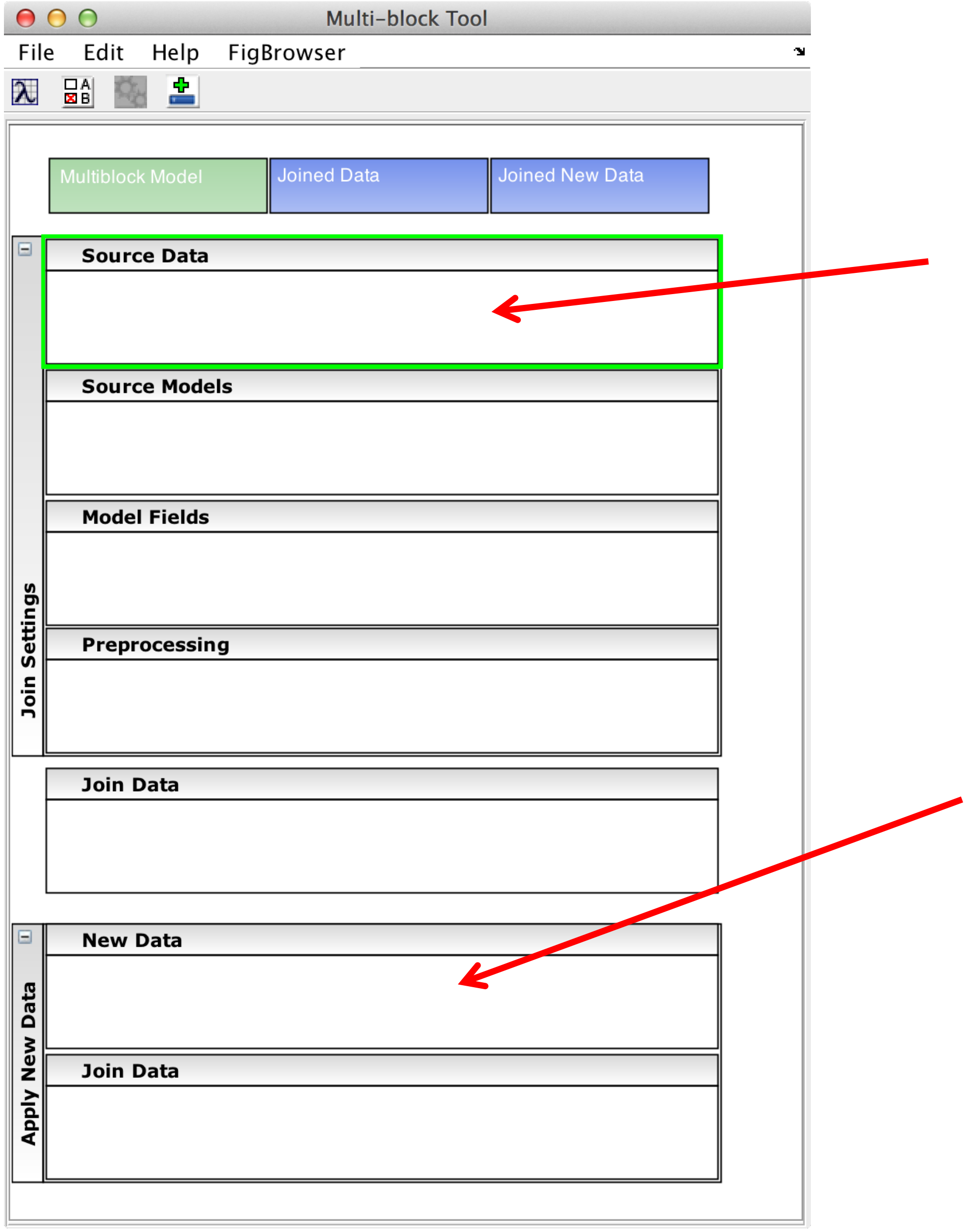
Sensitivity of MSPC Models

- Three experiments performed with 21 “induced” faults on:
  - TCP top power
  - RF bottom power
  - Cl2 flow
  - BCl3 flow
  - Chamber pressure
  - Helium chuck pressure
- Data available for Machine State, RF and OES
- Goal: Compare ability of models considered for detecting faults: best case and for routine data
- Generated realistic faults to test models

Example with Etch Data

- Available data: Machine, OES and RFM data for 104 normal wafers and 20 induced faults
- Data reduced just to mean over each batch

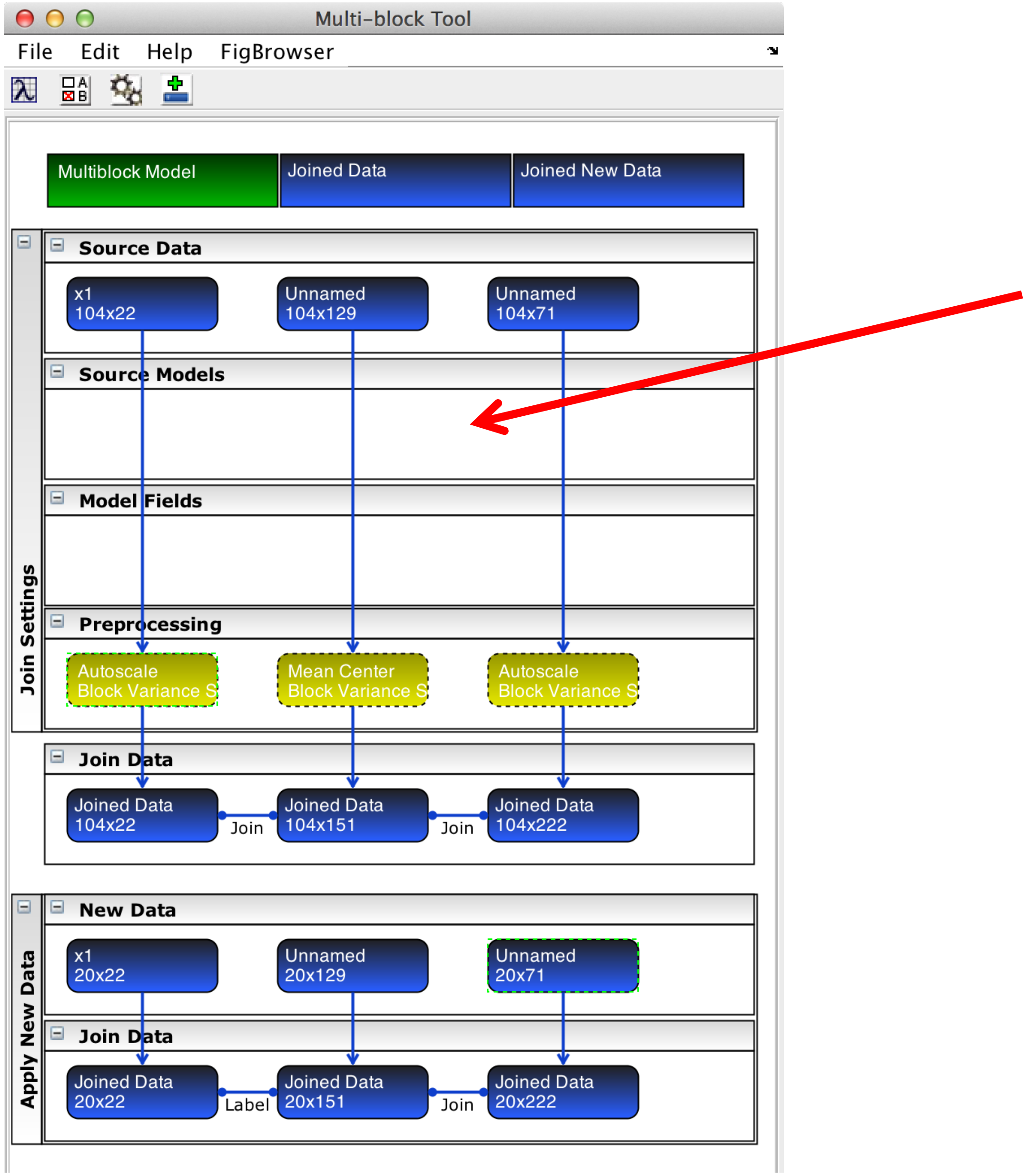
Multi-block Tool Interface



Drag calibration data sets here

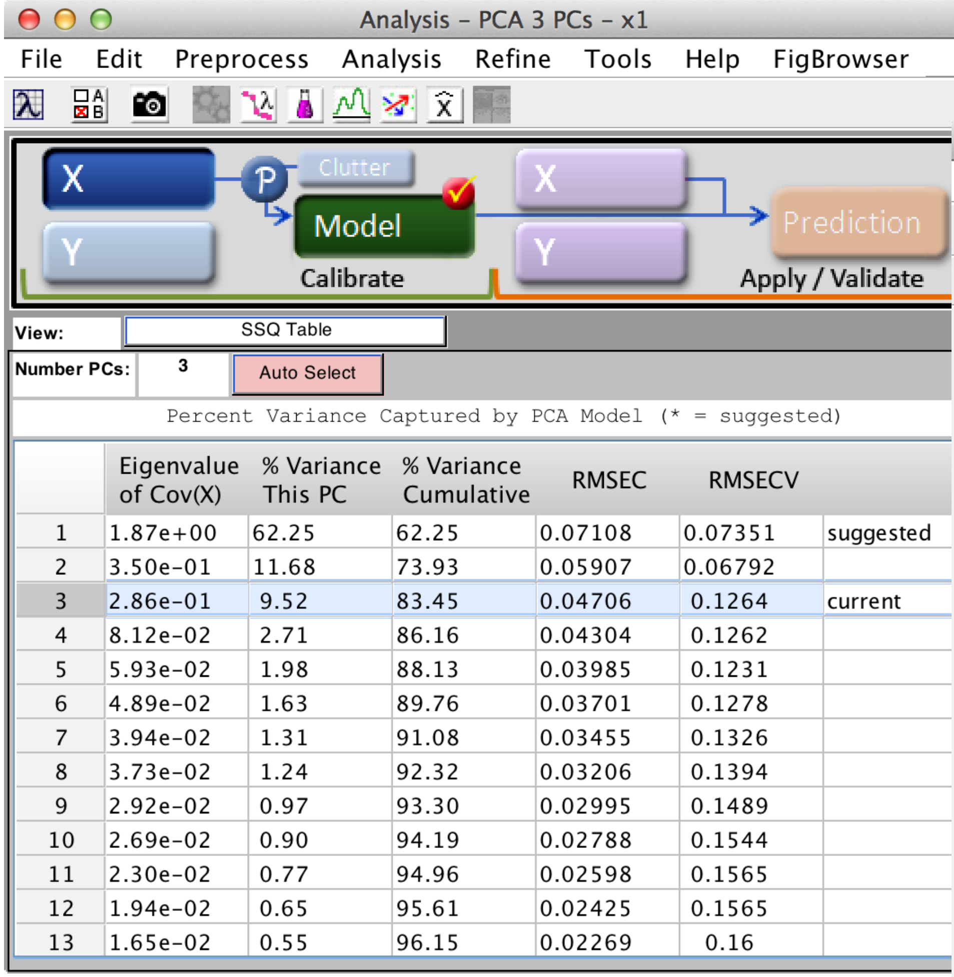
Drag test data sets here

Separately Preprocessed Then Joined Data



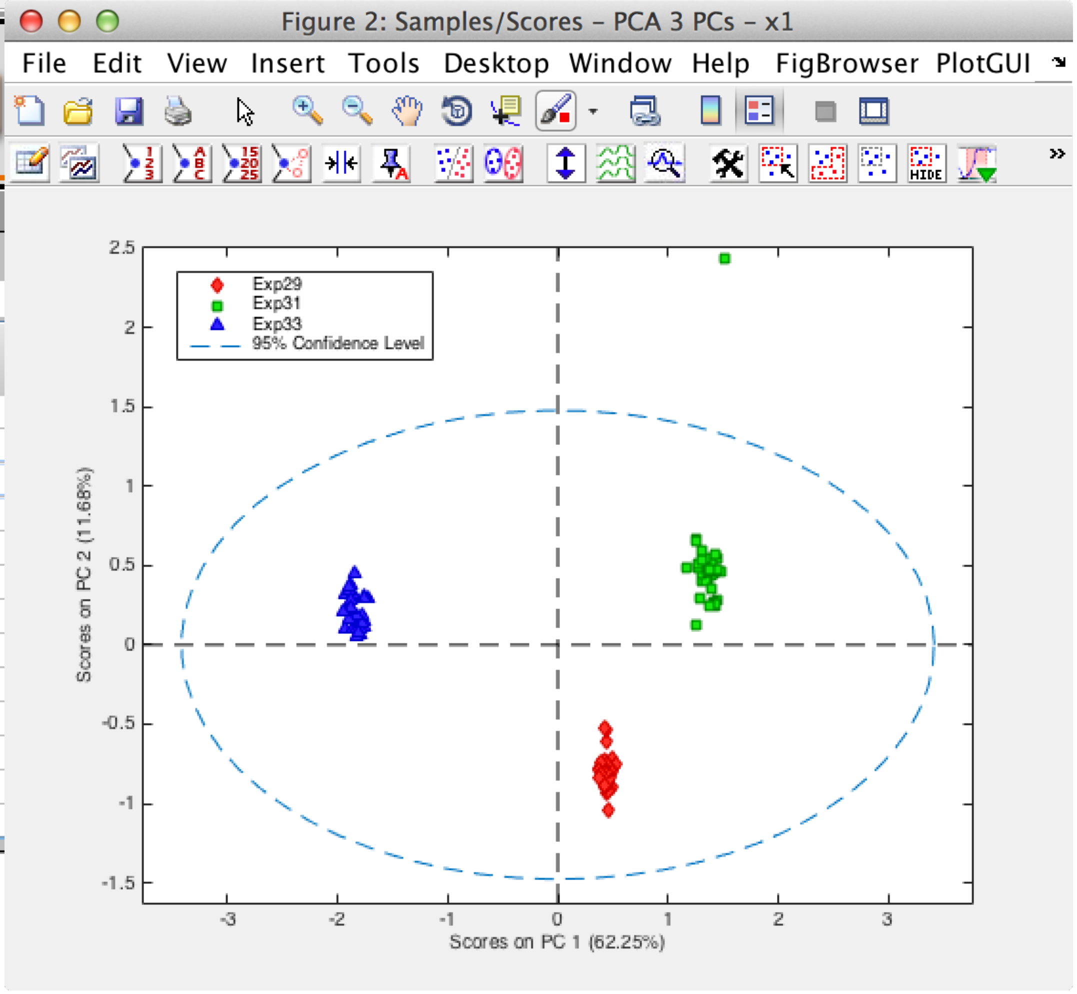
Or put models here for mid level fusion

Data pushed into PCA

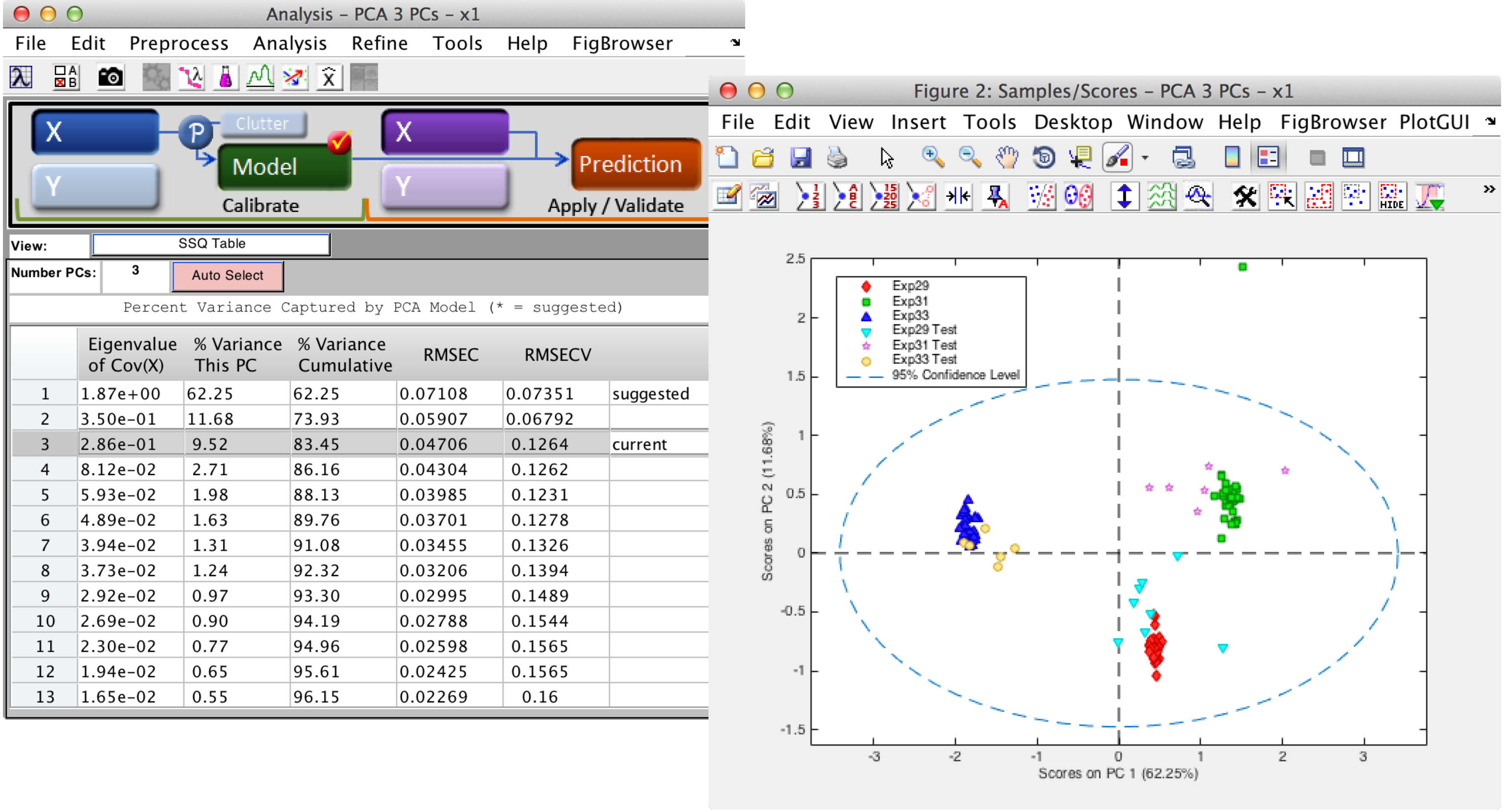


Percent Variance Captured by PCA Model (* = suggested)

	Eigenvalue of Cov(X)	% Variance This PC	% Variance Cumulative	RMSEC	RMSECV	
1	1.87e+00	62.25	62.25	0.07108	0.07351	suggested
2	3.50e-01	11.68	73.93	0.05907	0.06792	
3	2.86e-01	9.52	83.45	0.04706	0.1264	current
4	8.12e-02	2.71	86.16	0.04304	0.1262	
5	5.93e-02	1.98	88.13	0.03985	0.1231	
6	4.89e-02	1.63	89.76	0.03701	0.1278	
7	3.94e-02	1.31	91.08	0.03455	0.1326	
8	3.73e-02	1.24	92.32	0.03206	0.1394	
9	2.92e-02	0.97	93.30	0.02995	0.1489	
10	2.69e-02	0.90	94.19	0.02788	0.1544	
11	2.30e-02	0.77	94.96	0.02598	0.1565	
12	1.94e-02	0.65	95.61	0.02425	0.1565	
13	1.65e-02	0.55	96.15	0.02269	0.16	



With Test Data Loaded



Percent Variance Captured by PCA Model (* = suggested)

	Eigenvalue of Cov(X)	% Variance This PC	% Variance Cumulative	RMSEC	RMSECV	
1	1.87e+00	62.25	62.25	0.07108	0.07351	suggested
2	3.50e-01	11.68	73.93	0.05907	0.06792	
3	2.86e-01	9.52	83.45	0.04706	0.1264	current
4	8.12e-02	2.71	86.16	0.04304	0.1262	
5	5.93e-02	1.98	88.13	0.03985	0.1231	
6	4.89e-02	1.63	89.76	0.03701	0.1278	
7	3.94e-02	1.31	91.08	0.03455	0.1326	
8	3.73e-02	1.24	92.32	0.03206	0.1394	
9	2.92e-02	0.97	93.30	0.02995	0.1489	
10	2.69e-02	0.90	94.19	0.02788	0.1544	
11	2.30e-02	0.77	94.96	0.02598	0.1565	
12	1.94e-02	0.65	95.61	0.02425	0.1565	
13	1.65e-02	0.55	96.15	0.02269	0.16	

Redo at Mid-level

- Develop individual PCA models of data blocks
- Load models into Multi-block tool
- Choose model outputs
- Join and push into PCA
- Results similar

Conclusions I

- ASCA
  - for multi-set data typically from designed experiments
- MLASCA
  - for multi-level data typically from happenstance data (often semi-batch)
- ASCA and MLASCA allow new ways to partition and understand variance

Conclusions II

- Data Fusion methods combine multi-block data that share a common mode
- Data Fusion can be done at three levels
  - Low Level: joining blocks after preprocessing
  - Mid Level: joining model outputs such as scores
  - High Level: Combine predictions from multiple models in some sort of voting scheme
- Often brings out aspects of data that aren’t obvious in blocks analyzed separately

References

ASCA:

- Smilde, A.K., J.J. Jansen, H.C.J. Hoefsloot, R-J.A.N. Lamars, J. van der Greef, M.E. Timmerman, "ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data", Bioinformatics, 2005, 21, 3043-3048.
- Zwanenburg, G., H.C.J. Hoefsloot, J.A. Westerhuis, J.J. Jansen, and A.K. Smilde, "ANOVA-principal component analysis and ANOVA-simultaneous component analysis: a comparison". J. Chemometrics, 2011.

MLSCA:

- de Noord, O.E., and E.H. Theobald, Multilevel component analysis and multilevel PLS of chemical process data. J. Chemometrics 2005; 301–307
- Timmerman, M.E., Multilevel Component Analysis. Brit. J. Mathemat. Statist. Psychol. 2006, 59, 301-320.
- Jansen, J.J., H.C.J. Hoefsloot, J. van der Greef, M.E. Timmerman and A.K. Smilde, Multilevel component analysis of time-resolved metabolic fingerprinting data. Analytica Chimica Acta, 530, (2005), 173–183.