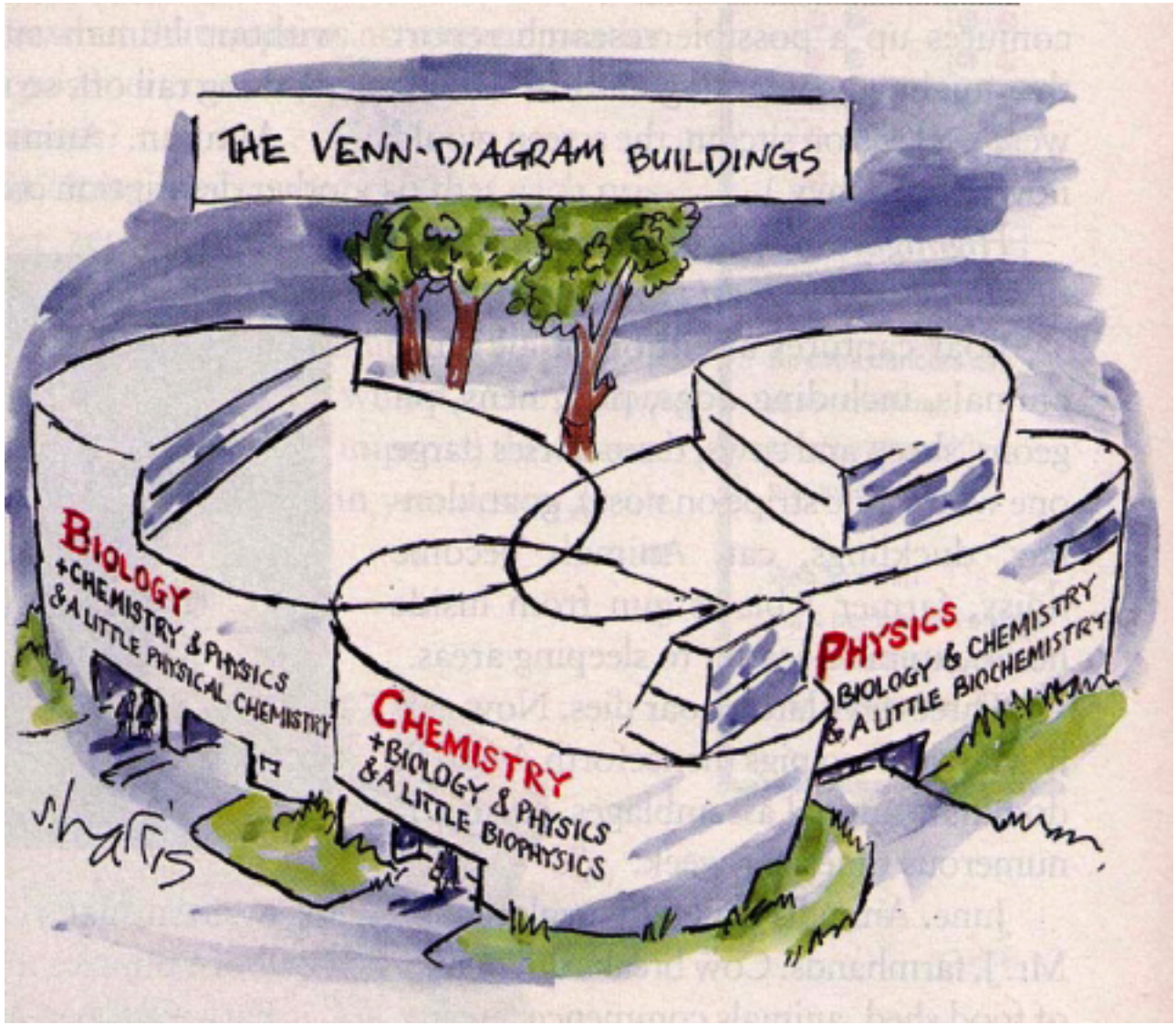
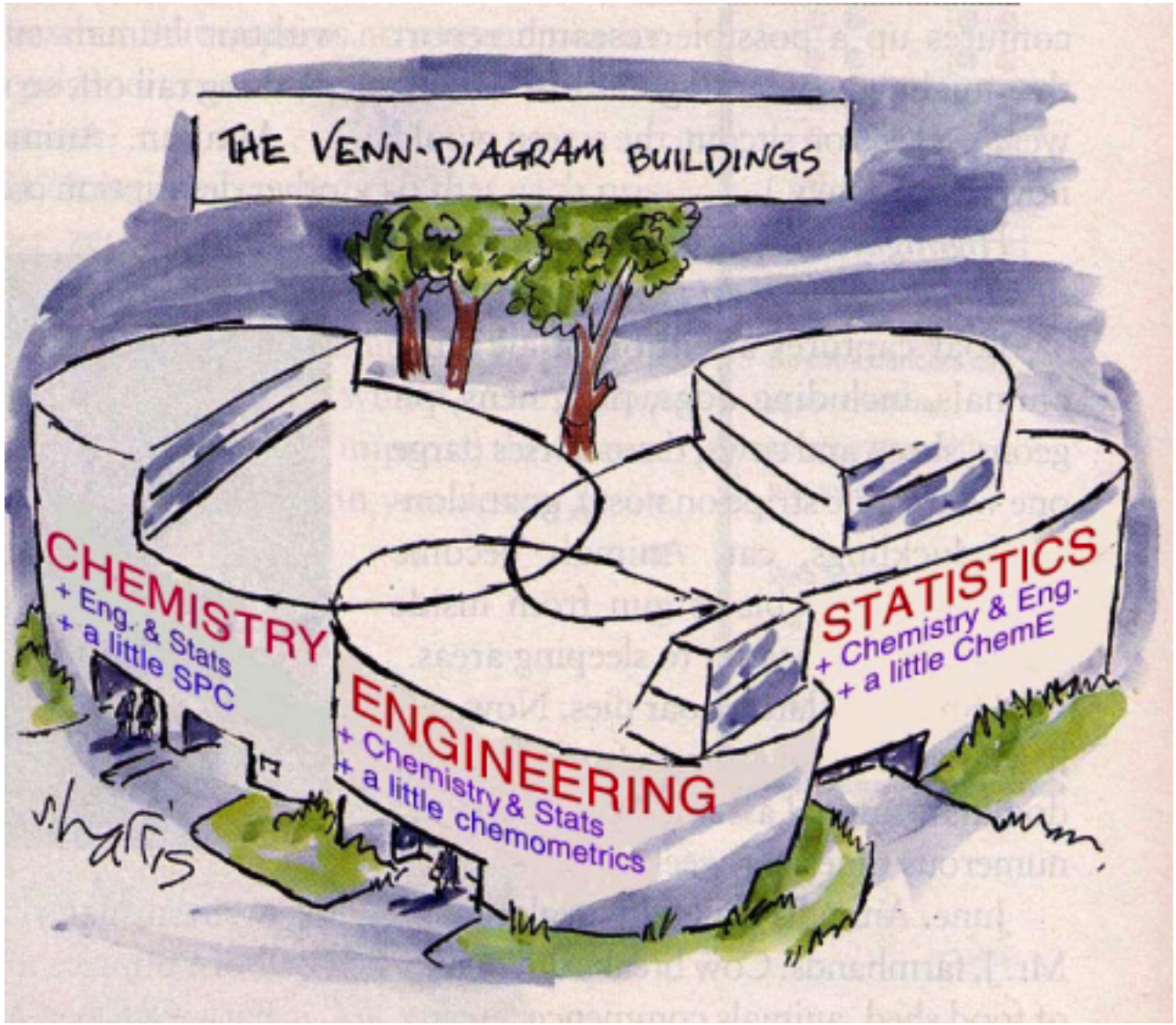
Getting to Multiway: A Roadmap for Batch Process Data

Barry M. Wise, Jeremy M. Shaver and Neal B. Gallagher
Eigenvector Research, Inc.
Wenatchee, WA, USA

EIGENVECTOR RESEARCH INCORPORATED



EIGENVECTOR RESEARCH INCORPORATED



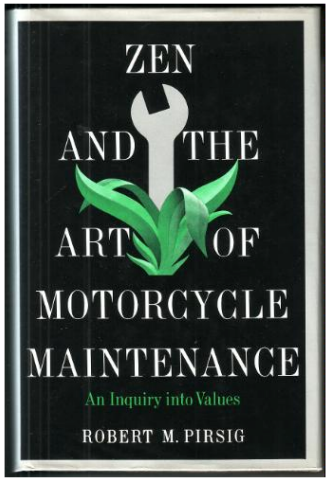
EIGENVECTOR RESEARCH INCORPORATED

Outline

- General principles of SPC and fault detection
- Batch Chemical Processes
- Roadblocks
- Models
- The Roadmap
- Example: Comparison on Dupont Data
- Conclusions

General Principles of Fault Detection

- Process monitoring / Fault Detection / Statistical Process Control / Multivariate SPC / Batch SPC...
- Methods rely on a model that describes normal and/or desirable* operation
  - New data compared with model of "normal" data
  - Often much is learned from this model and the process of creating it!
  - Data considered normal to the process is likely not the same data useful for constructing regression models.



* Quality isn't JUST what you like, it IS what you like.
-Robert M. Pirsig, "Zen and the Art of Motorcycle Maintenance," William Morrow & Co. 1974

EIGENVECTOR RESEARCH INCORPORATED

Different Modeling Approaches

- Theoretical
  - Mathematical models, constructed from first principles
  - Applicable to information-sparse systems
    - good given satisfactory models and sufficient sensors
    - often time consuming to develop models
    - difficult to apply to large scale systems
- Empirical
  - Derived directly from process data
  - Applicable to data-rich systems
    - requires some redundancy in the data (fewer states than measurements)
    - highly dependent upon the quantity, quality and reliability of process instruments

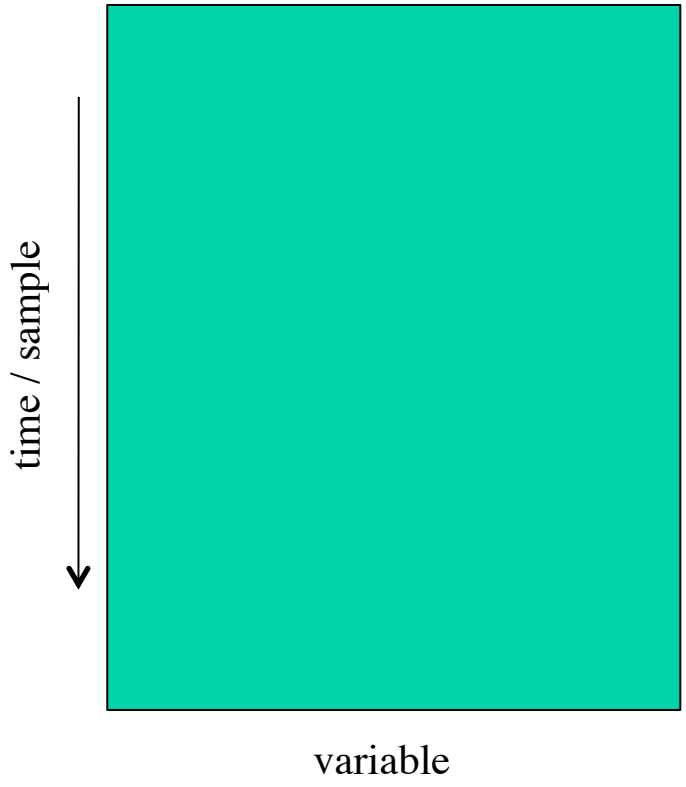
General Principles of Fault Detection

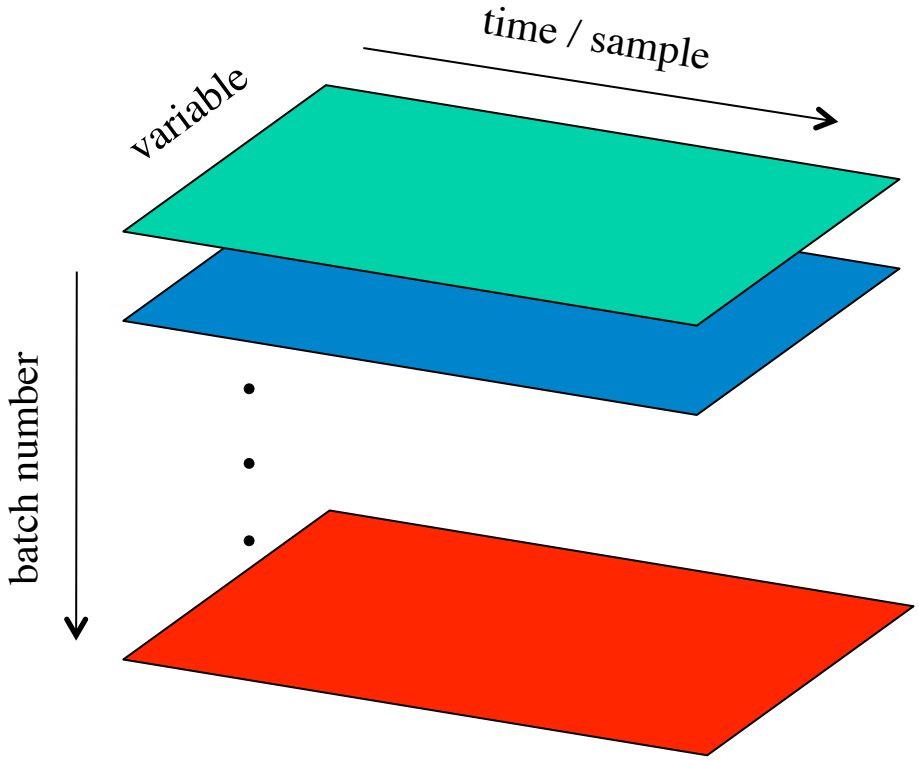
- Operating data is compared to the process model to determine if the process condition is nominal
  - do the new measurements look like the old ones or are they significantly different?
  - is the process in control?
- If different, it is useful to have diagnostic information about how it may be different
  - why is the process out of control?
  - some models are very good at providing diagnostics

Batch Chemical (and Manufacturing) Processes

- Many things made in batch (as opposed to continuous) processes:
  - Pharmaceuticals, enzymes
  - Food (cheese, yogurt), beverages (beer, wine)
  - Semiconductors
  - Polymers
  - etc....
- Batch data nominally 3-way
  - Process measurements (sensors, spectroscopy)
  - Batch running time
  - Batch number

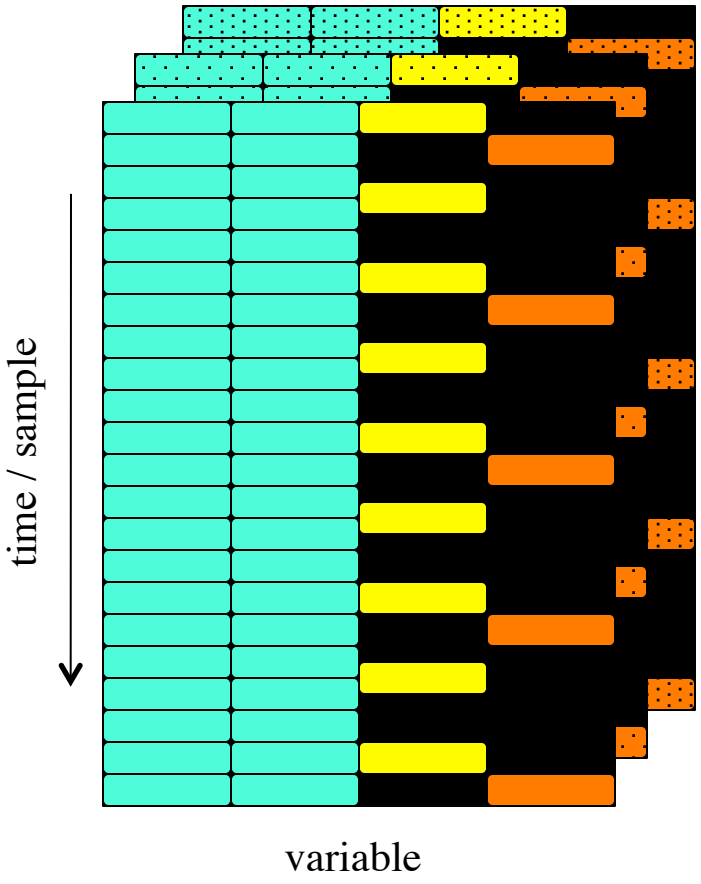
Batch Data



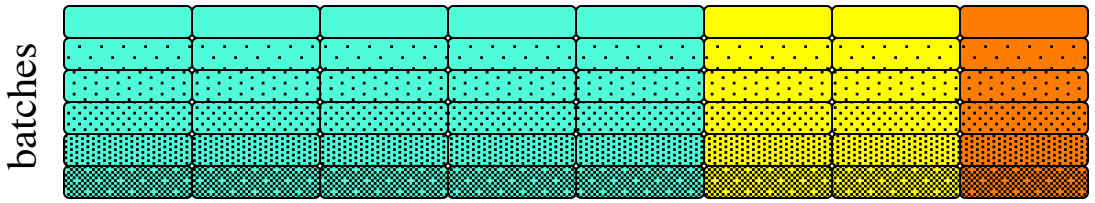


EIGENVECTOR RESEARCH INCORPORATED

Different sample rates



- How to handle?
  - Zero order hold
  - Interpolate
  - Treat as missing?
- Doesn't matter in MPCA as long as it is consistent!



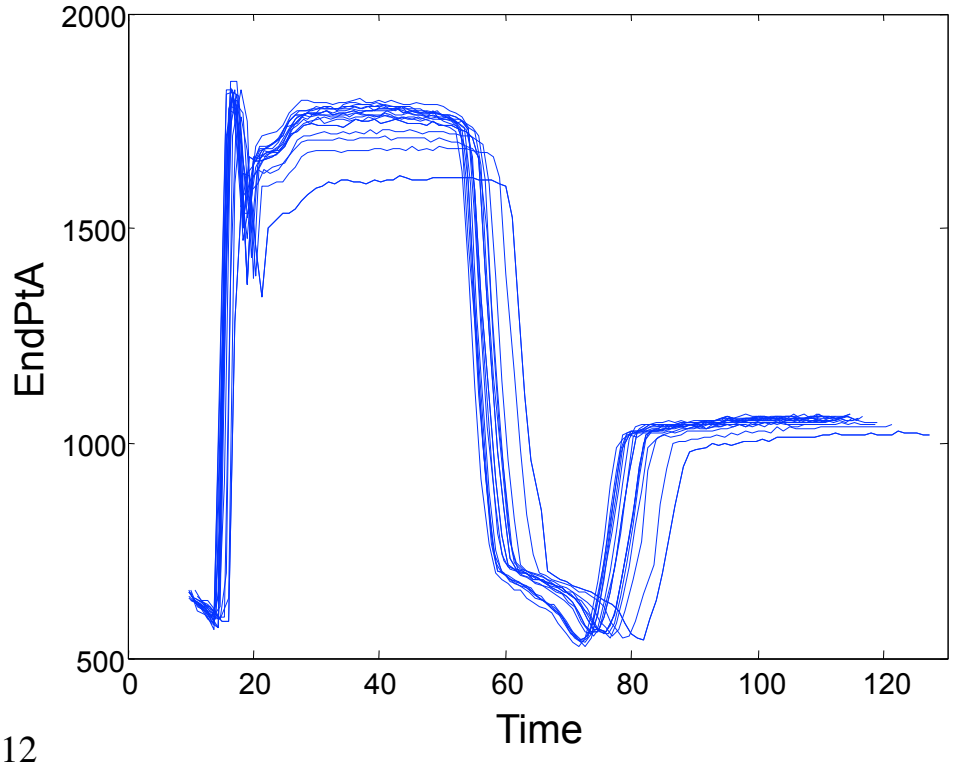
EIGENVECTOR RESEARCH INCORPORATED

Batch Process Monitoring Data Problems

- The objective is to monitor batch-to-batch, but
  - data can be messy
    - typically includes start-up and shut-down phases that are not of interest
    - might be interesting if monitoring controller performance
  - periods of "steady-state" where not much is changing
  - variable record lengths
  - lots of data!
- Reduce to a set of more compact descriptors?
  - show an example…

Process Data Alignment and Dilation

- Batches "mature" at different rates
- Leads to files of different length
- Important transitions occur at different times

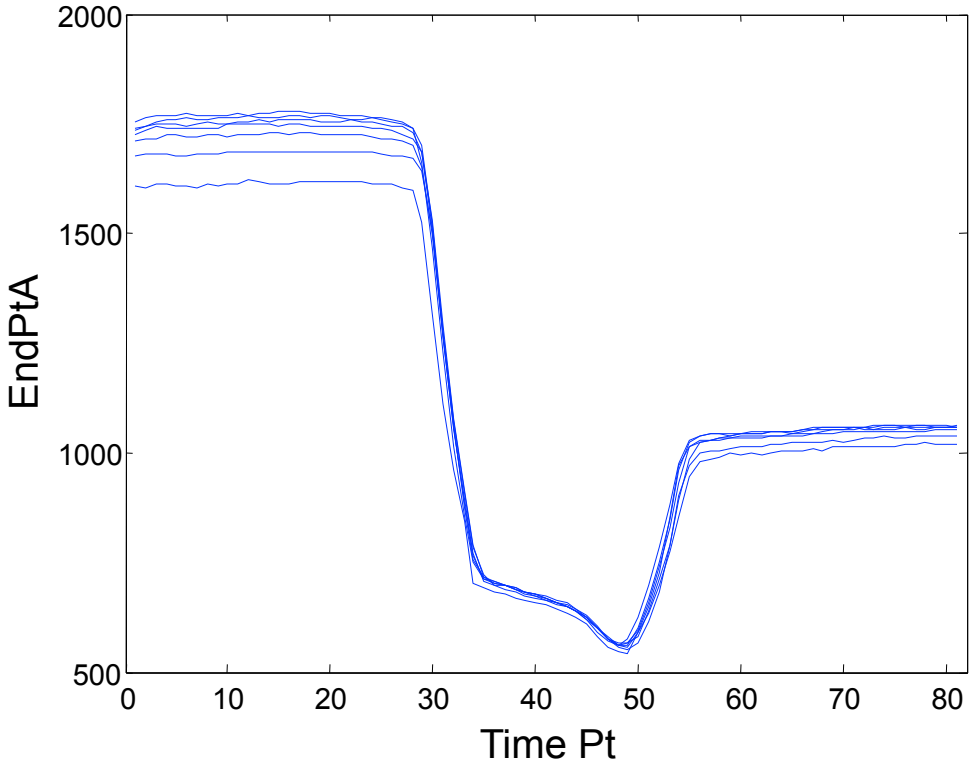


- Misalignment adds rank irrelevant to process monitoring.
  - model must account for time shifts in the process data.
  - Irrelevant variance often results in a reduction of model sensitivity.

EIGENVECTOR RESEARCH INCORPORATED

Aligned Process Data

- Many ways to align and/or warp
  - Align and truncate
  - Correlation Optimized Warping (COW)
  - Dynamic Time Warping (DTW)



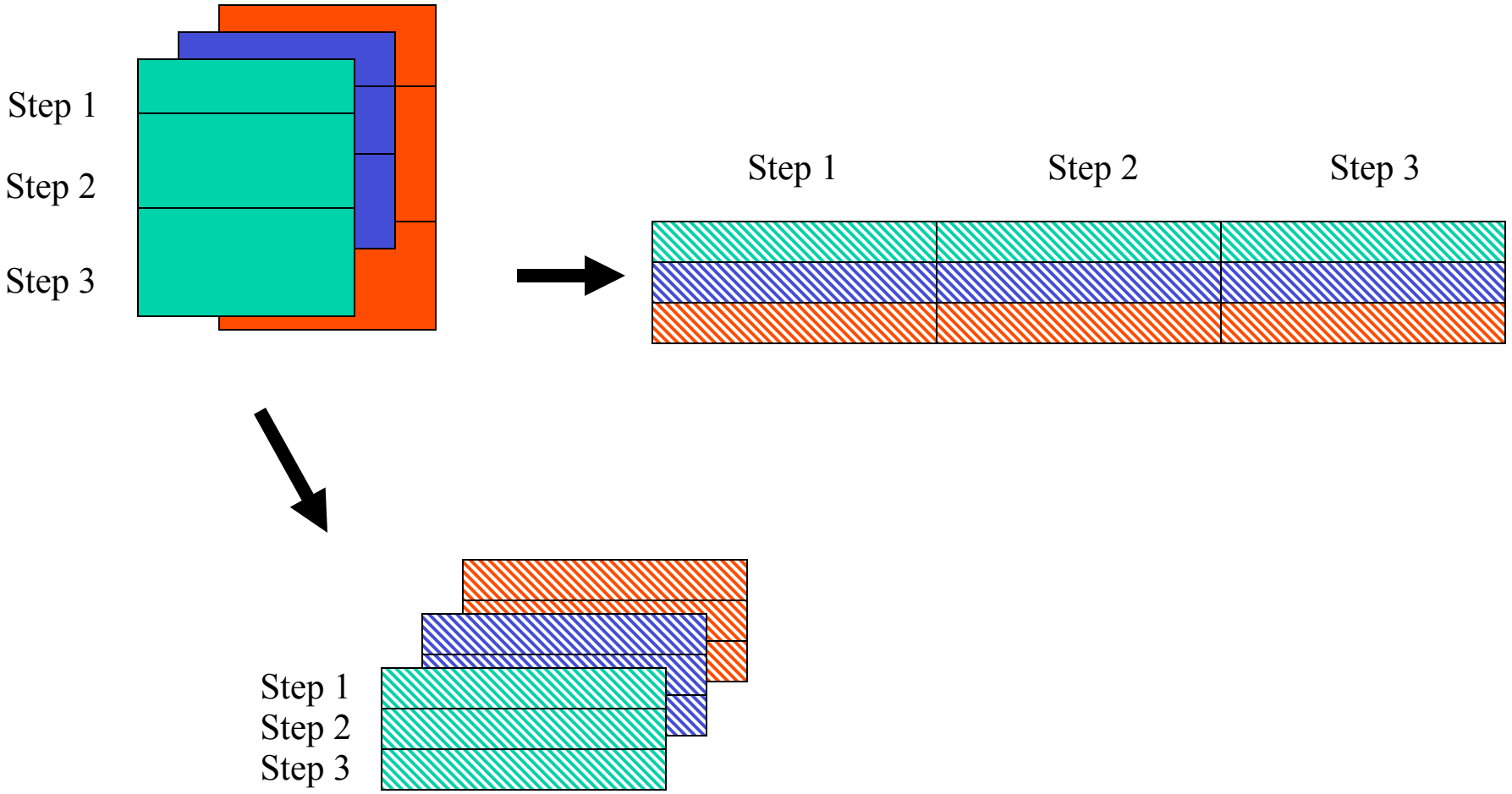
Alternative:
Summarize data
over process steps

EIGENVECTOR RESEARCH INCORPORATED

Data Summary Approach

- Convert data into alternate set of descriptors
- If process has multiple steps, calculate parameters that describe each step
  - mean
  - standard deviation
  - slope
  - length (time) of the step
  - etc....

Summary Variables



EIGENVECTOR RESEARCH INCORPORATED

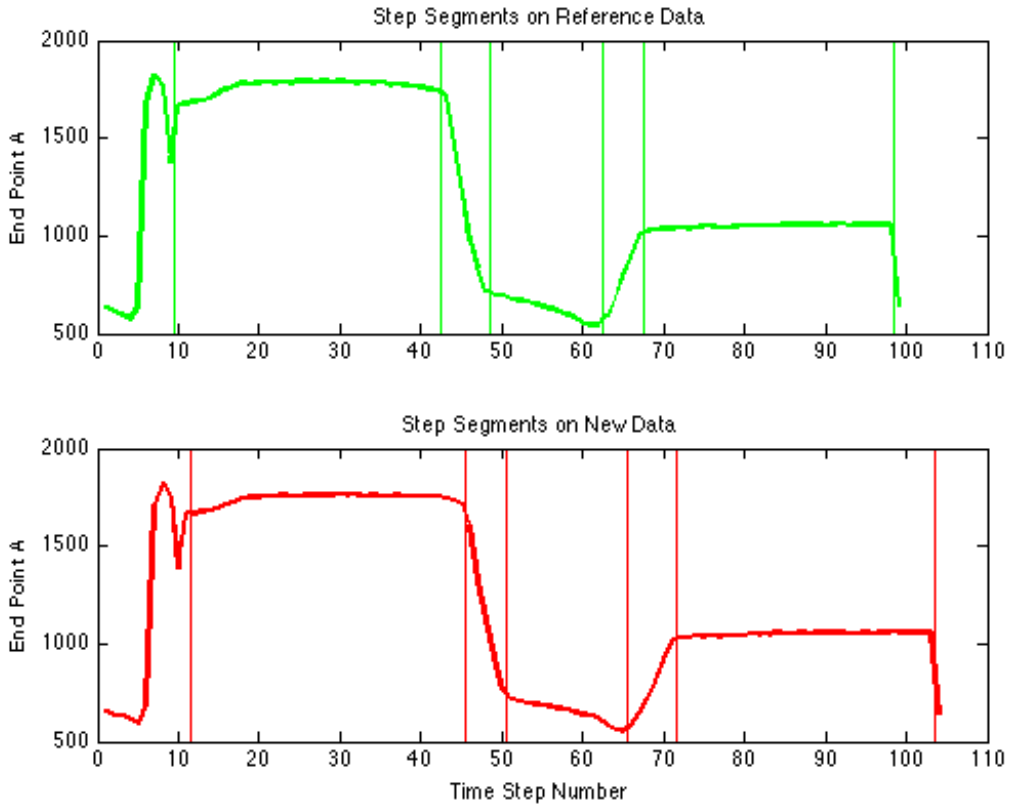
Summary Variables

- Pros
  - Conceptually simple
  - Some time information retained
  - Noise reduction
  - Reduces number of variables (vs. MPCA)
- Cons
  - Further from original data
  - May not have step numbers to work with

Creation of Pseudo-steps

- Several ways to do this
  - Manual assignment followed by warping
    - Break reference process variable into “sensible” segments (manually)
    - Assign step numbers
    - Warp new data onto reference
    - "Reverse warp" reference step numbers into new data
  - Automated peak picking
    - take first or second derivative of reference process variable
    - use peak peaking algorithm to find transitions

Example of Step Creation



EIGENVECTOR RESEARCH INCORPORATED

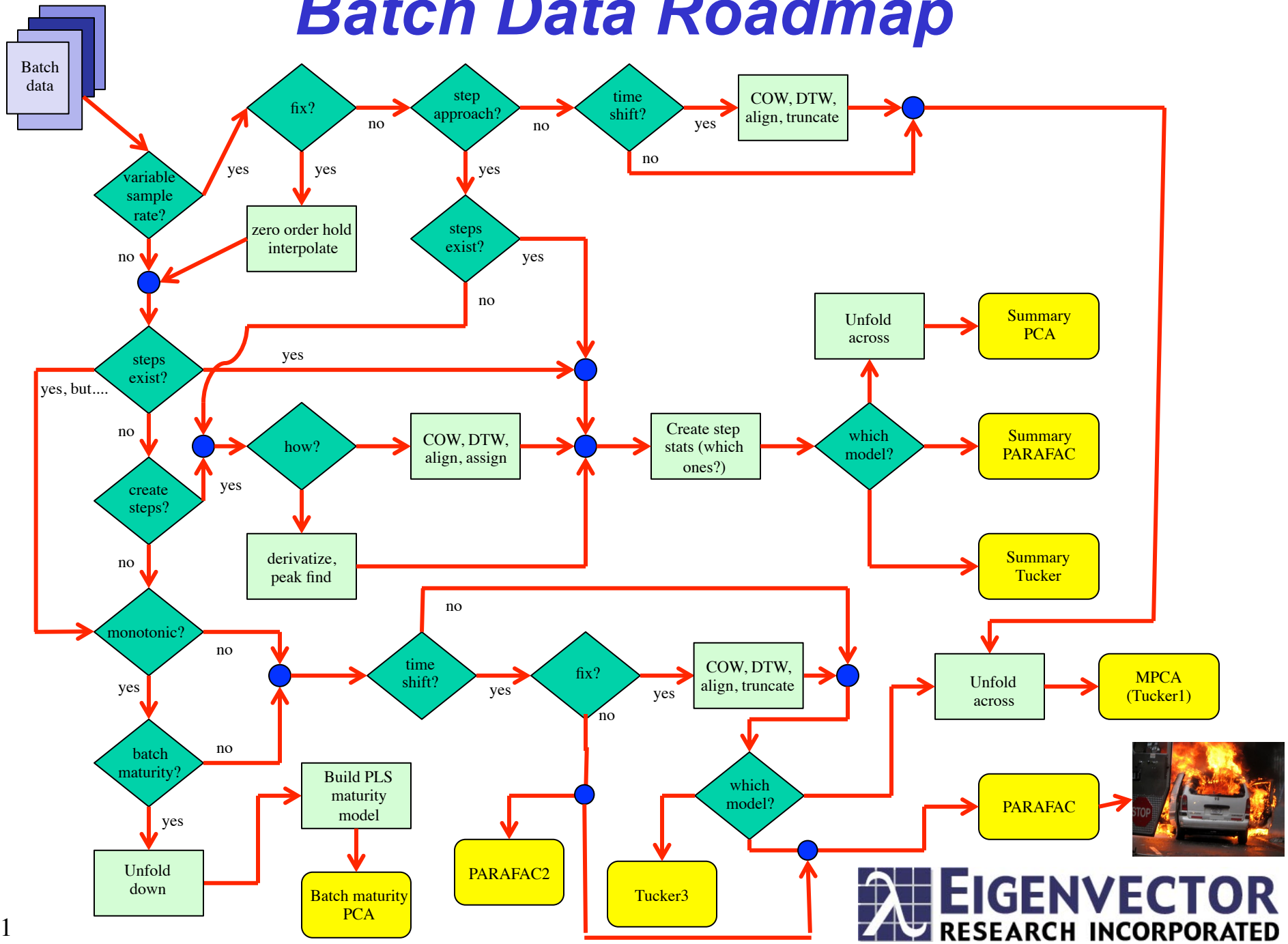
Batch Maturity Models

- Build PLS (or other) model that predicts extent of reaction or “batch time” for each time point
- Use model to predict where points should be on the time axis
  - can use as basis for warping, then can use PARAFAC, or MPCA model
  - use conventional PCA model (unfold down), make limits on scores, residuals etc. be a function of batch time

Which Model?

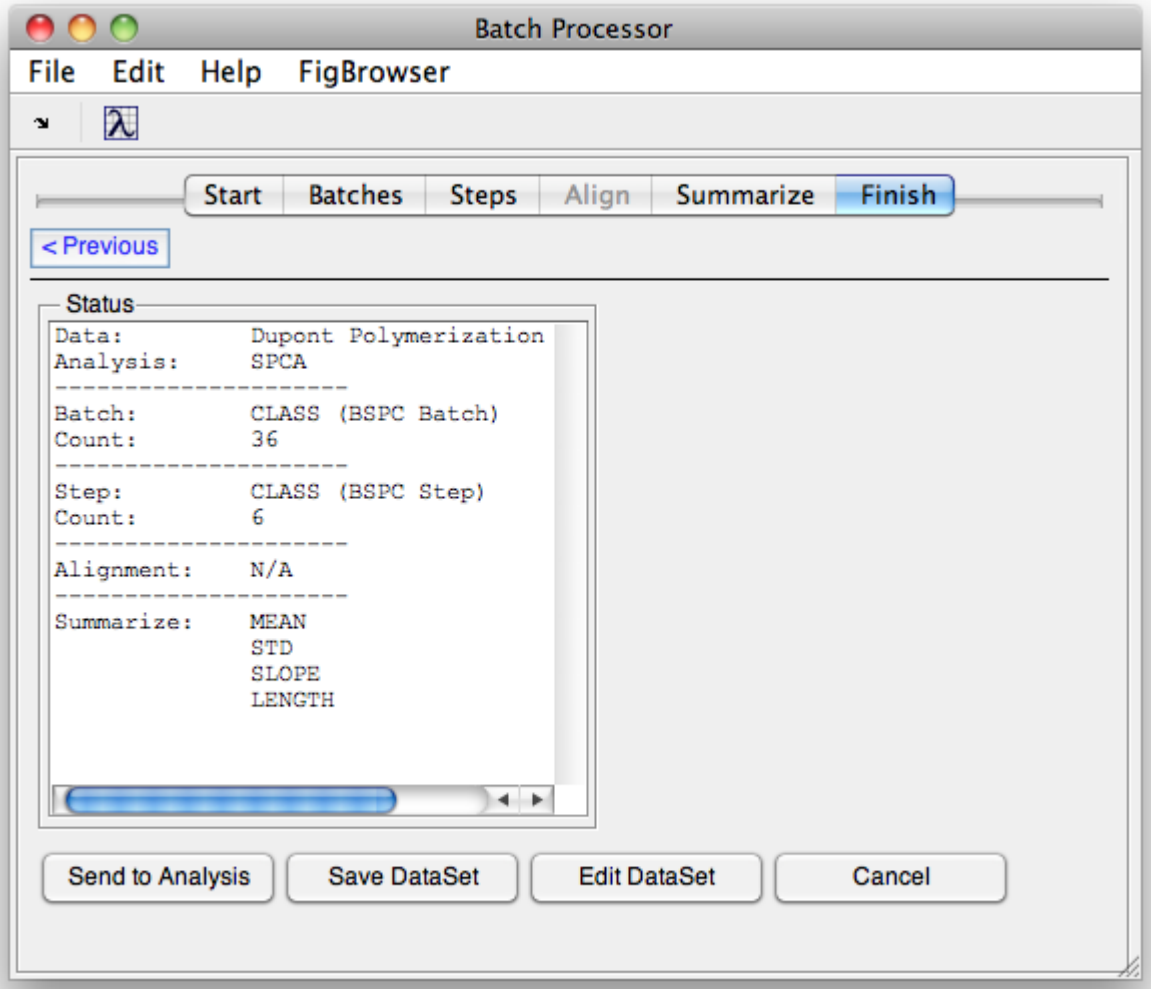
- Many to choose from!
  - MPCA (aka Tucker-1)
  - Summary PCA (MPCA on summary variables)
  - PARAFAC
  - Summary PARAFAC
  - PARAFAC-2
  - Tucker-3
  - Summary Tucker-3
  - Batch Maturity PCA
  - Tucker-2?

Batch Data Roadmap



EIGENVECTOR RESEARCH INCORPORATED

Batch Processor Tool



EIGENVECTOR RESEARCH INCORPORATED

Dupont Batch Data

- 10 process variables (sanitized)
  - 100 time intervals each
  - TempR 1, TempR 2, TempR 3, Press 1, Flow 1, TempC 1, TempC 2, Press 2, Press 3, Flow 2
- Calibration: 1 to 36 (normal batches)
- Test: 37 to 55 (one normal and seven faults)
  - Batches 40, 41, 42, 50, 51, 53, 54 and 55 had the final quality measurement well outside the acceptable limit
  - Batches 38, 45, 46, 49 and 52 were above or very close to that limit.
  - Batches 38, 40, 41 and 42 cannot be identified as abnormal batches.*
  - Additional batches 37, 39, 43, 44, 47, 47 and 48 were identified as somewhat unusual and were not included in the calibration set.*
- Described in

*Nomikos, P. and J.F. MacGregor, "Multivariate SPC charts for monitoring batch processes," Technometrics, **37**(1), 1995.

Methods

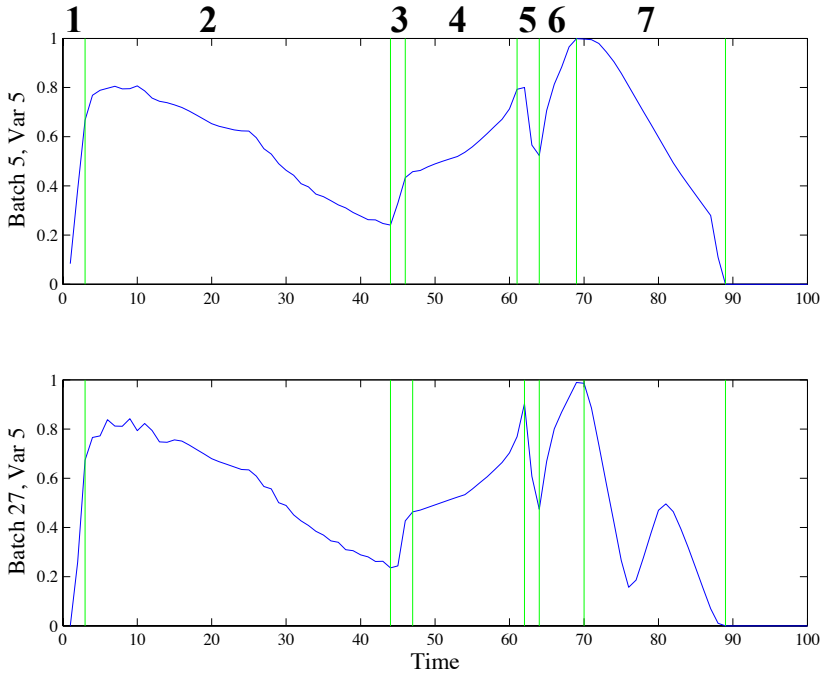
- Multi-Way PCA (MPCA, Tucker1)
  - block/group scaling on raw data
  - after COW
- Summary PCA (SPCA)
- Summary PARAFAC
- PARAFAC
- Batch Maturity PCA

SPCA

- PCA for summary variables
  - used steps 2-7
  - summarized by mean and step length
  - total variables = 10 * 6 + 6 = 66

Variable 5 is a feed flow and can be used to identify steps

Steps for two example batches



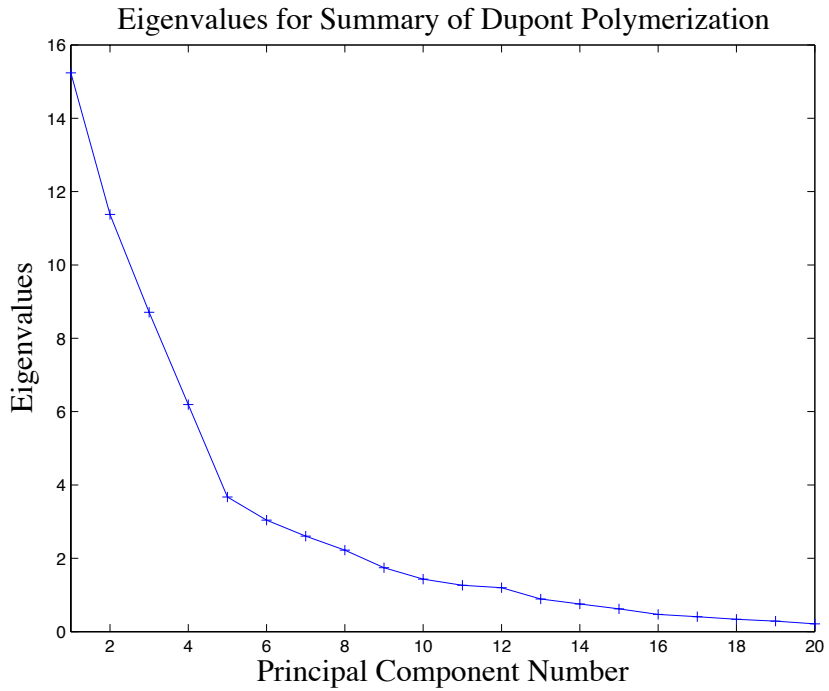
EIGENVECTOR RESEARCH INCORPORATED

SPCA Variance Captured

```
X-block: Summary of Dupont Polymerization 36 by 64
Included: [ 1-36 ] [ 1-53 55-64 66 ]
Preprocessing: Autoscale
Num. PCs: 4
```

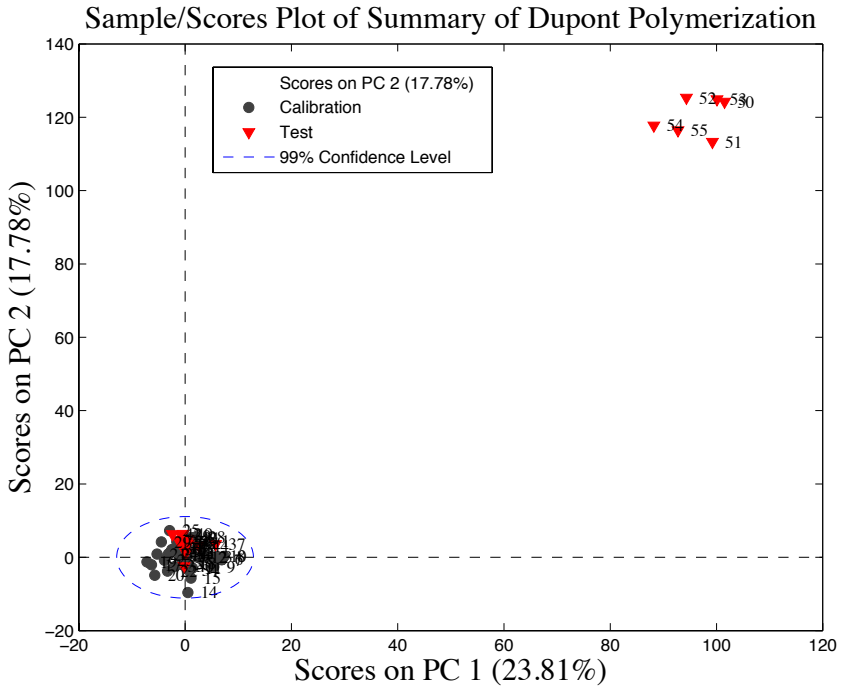
Percent Variance Captured by PCA Model

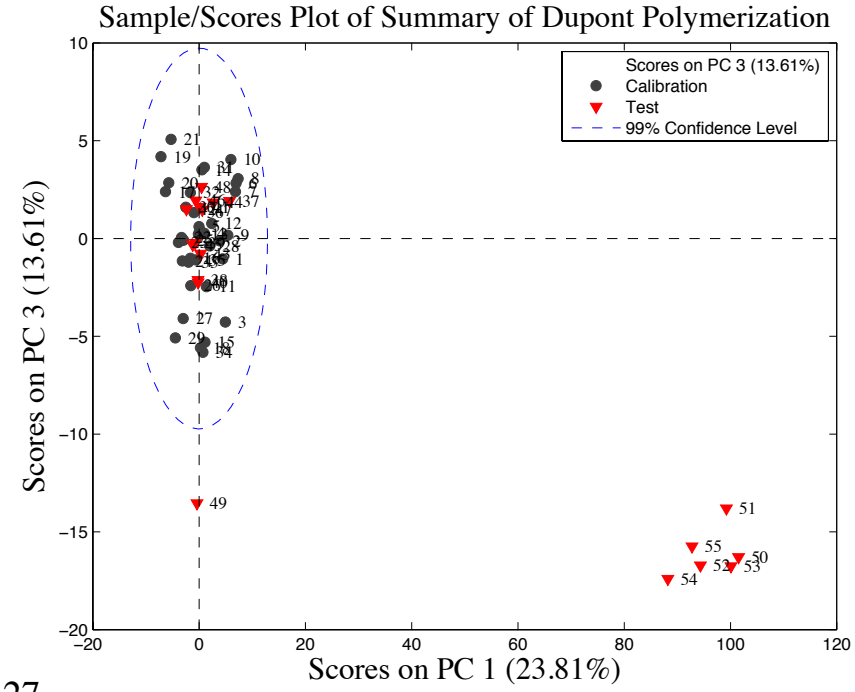
Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	15.2	23.81	23.81
2	11.4	17.78	41.59
3	8.71	13.61	55.20
4	6.19	9.67	64.87

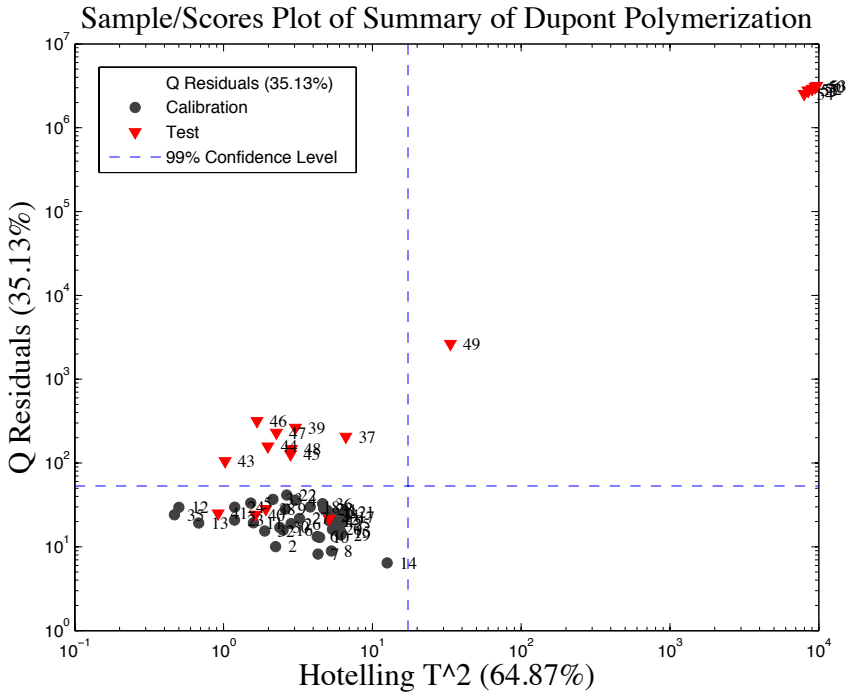


EIGENVECTOR RESEARCH INCORPORATED

SPCA Scores



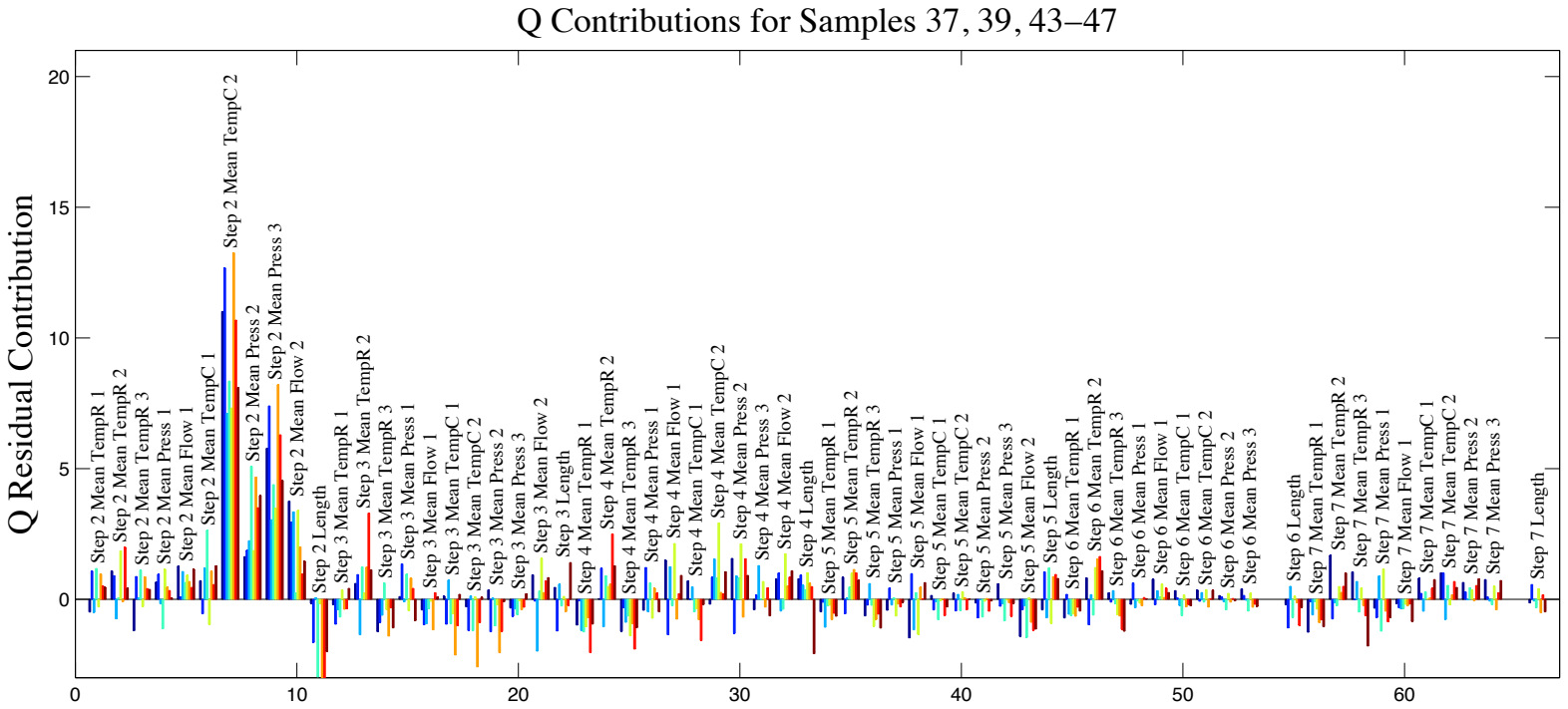




Out on both: 49, 50-55
Out on Q only: 37, 39, 43-48
Total coefficients: 4x66 = 264

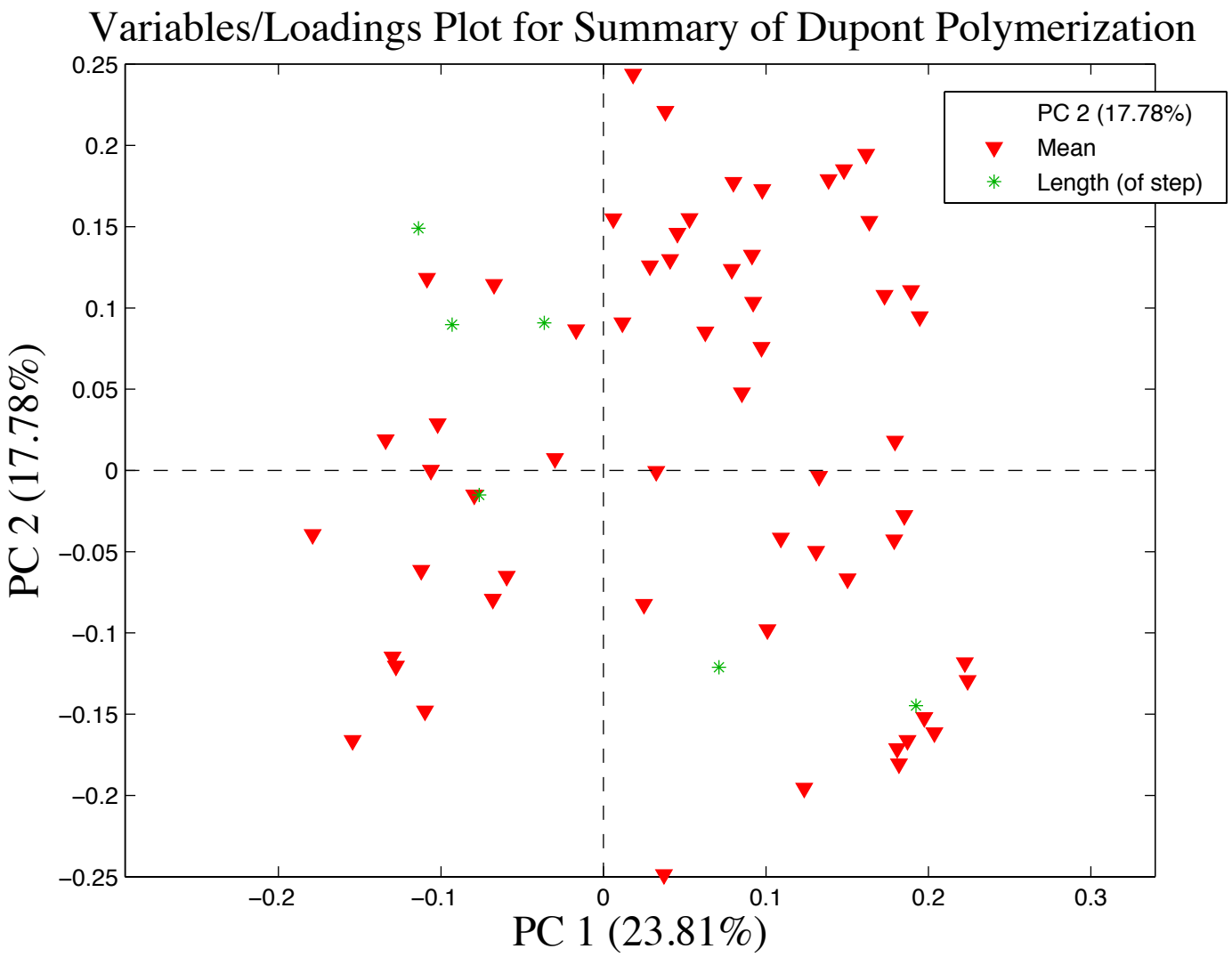
EIGENVECTOR RESEARCH INCORPORATED

Varcap, Q Contributions



EIGENVECTOR RESEARCH INCORPORATED

Loadings



EIGENVECTOR RESEARCH INCORPORATED

MPCA on Original Data

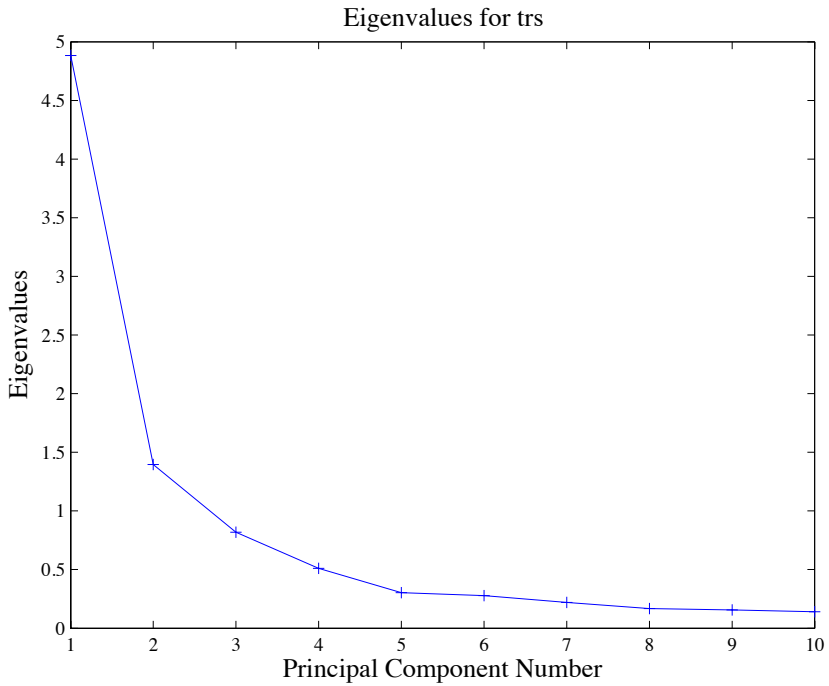
- MPCA with block scaling
- 100 time steps x 10 variables = 1000
- scale each block of the 100 new variables corresponding to individual original variables to unit variance and zero mean

MPCA Variance Captured

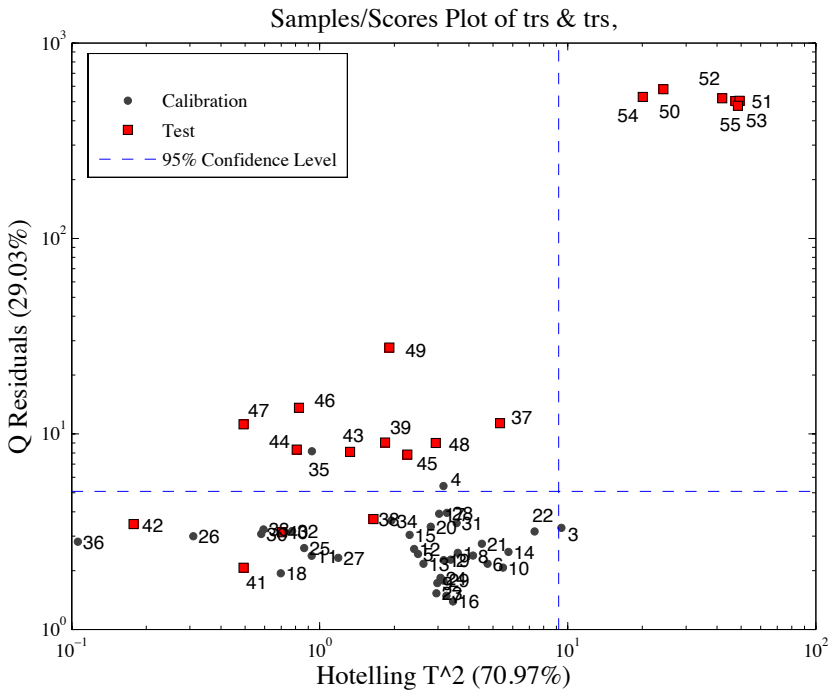
X-block: 47 by 1000
Preprocessing: Groupscale
Num. PCs: 3

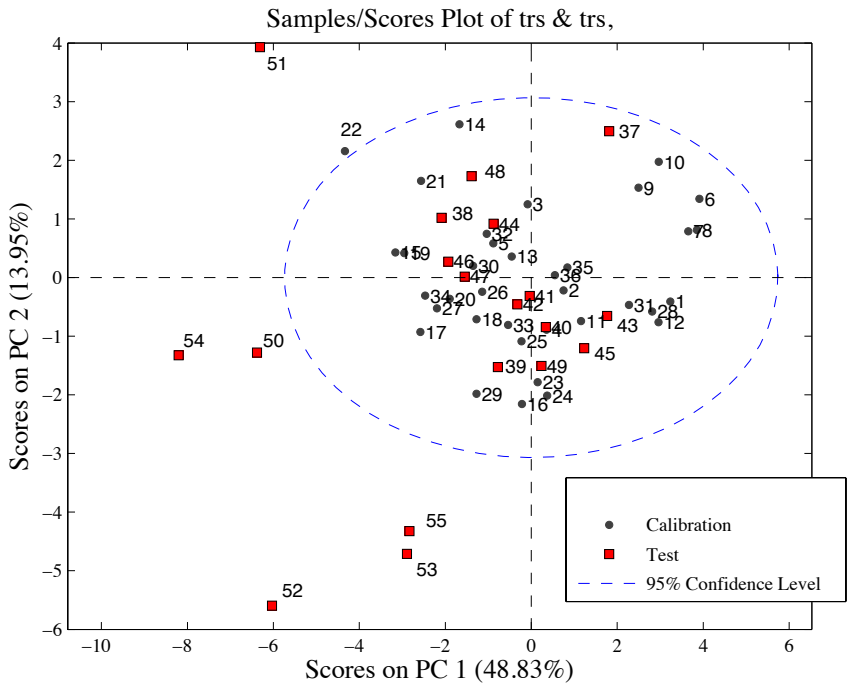
Percent Variance Captured by PCA Model

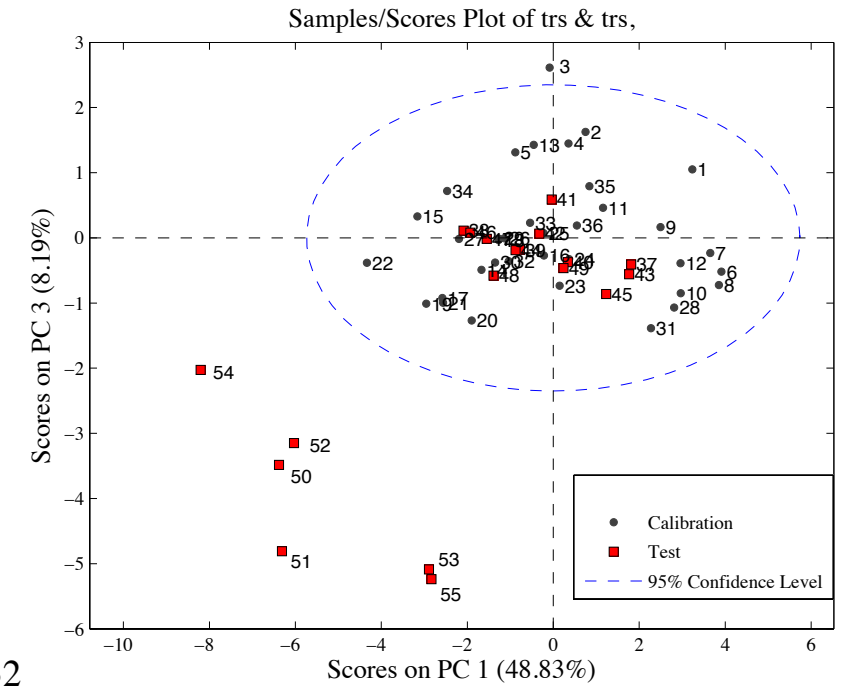
Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	4.88e+00	48.83	48.83
2	1.40e+00	13.95	62.78
3	8.19e-01	8.19	70.97



EIGENVECTOR RESEARCH INCORPORATED



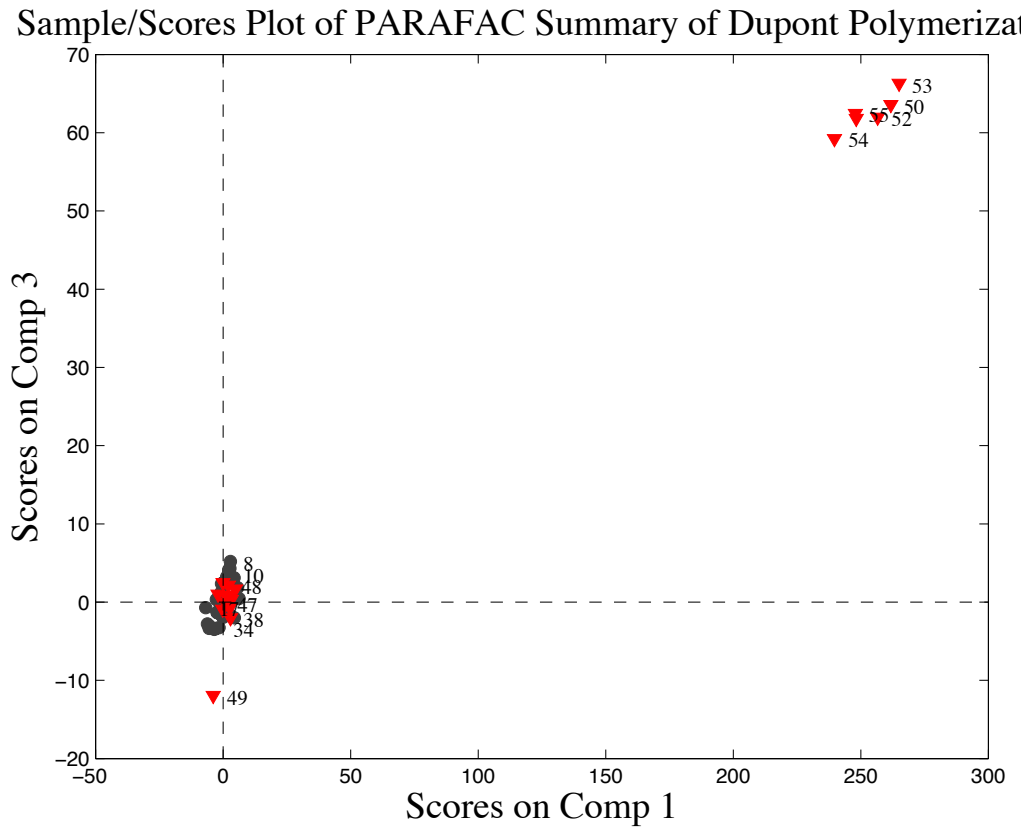




Out on both: 50-55
Out on Q only: 37, 39, 43-49
Total coefficients: 3x1000 = 3000!

EIGENVECTOR RESEARCH INCORPORATED

Summary PARAFAC

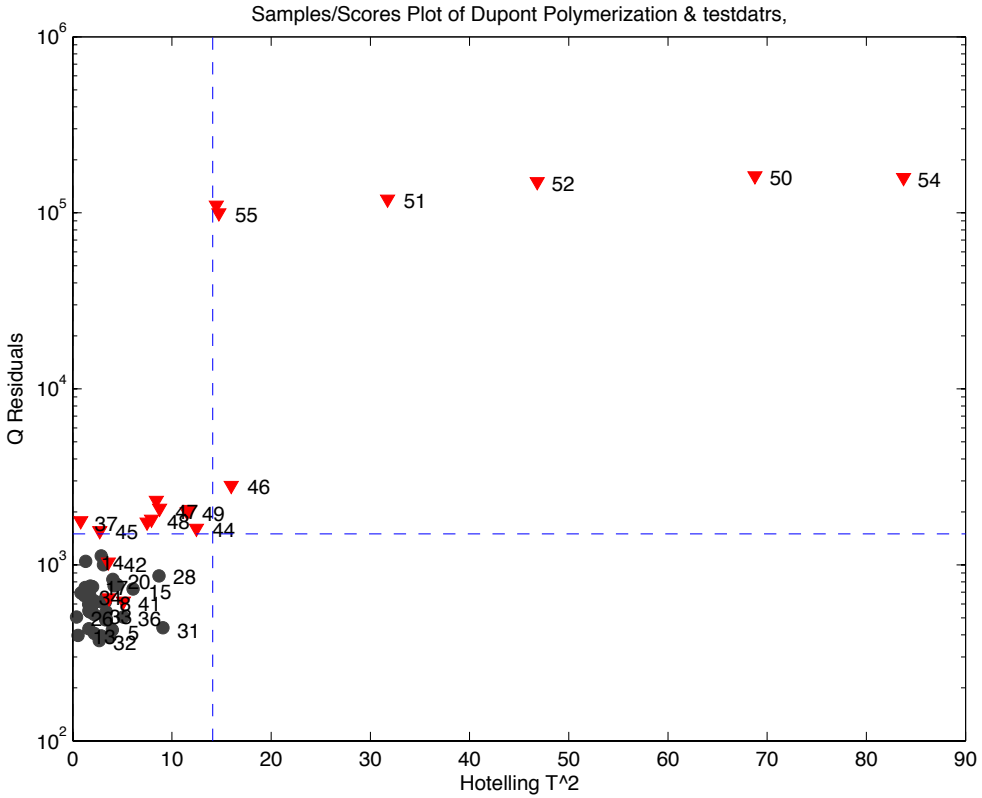


Out on both: 49-55
Out on Q only: no additional
Total coefficients:
3x11 + 3*6 = 51

Not as sensitive as SPCA and MPCA, but far fewer parameters

EIGENVECTOR RESEARCH INCORPORATED

PARAFAC on Original Data



Out on both: 46, 50-55
Out on Q only:37, 39, 43-45, 47-49

Total coefficients:
3x10 + 3*100 = 330

Virtually same results as SPCA and MPCA

EIGENVECTOR RESEARCH INCORPORATED

Issues?

- Plenty!
- MPCA models easy to overfit, PARAFAC models sometimes not flexible enough
- Have not addressed run time application of models to partial batches
  - not so tough IF warping or step creation isn't an issue...
  - but hard to warp partial batches
  - some models can't be fit to partial data records (PARAFAC2)

Conclusions

- Too many options!
- Hard to know what method is best for particular application
- Challenging to implement in software that mere mortals can use
- If multi-way methods are to be adopted for batch processes, needs to be streamlined