Monitoring and Fault Detection with Multivariate Statistical Process Control (MSPC) in Continuous and Batch Processes

Barry M. Wise, Ph.D.
Eigenvector Research, Inc.
Manson, WA

Outline

- Definition of Chemometrics
- Favorite tools
  - Principal Components Analysis (PCA)
  - Partial Least Squares Regression (PLS)
  - Multi-way methods
- Opportunities in PAT
  - Multivariate Statistical Process Control (MSPC)
  - Image analysis on tablets
  - Predicting monitored or controlled variables
  - Batch MSPC

Chemometrics

Chemometrics is the chemical discipline that uses mathematical and statistical methods to

1) relate *measurements* made on a *chemical* system to the *state* of the system, and
2) design or select optimal *measurement* procedures and experiments.

Multivariate Analysis

Multivariate Statistical Analysis is concerned with data that consists of *multiple measurements* on a number of individuals, objects, or data samples.

The measurement and analysis of *dependence between variables* is fundamental to multivariate analysis.

EIGENVECTOR RESEARCH INCORPORATED

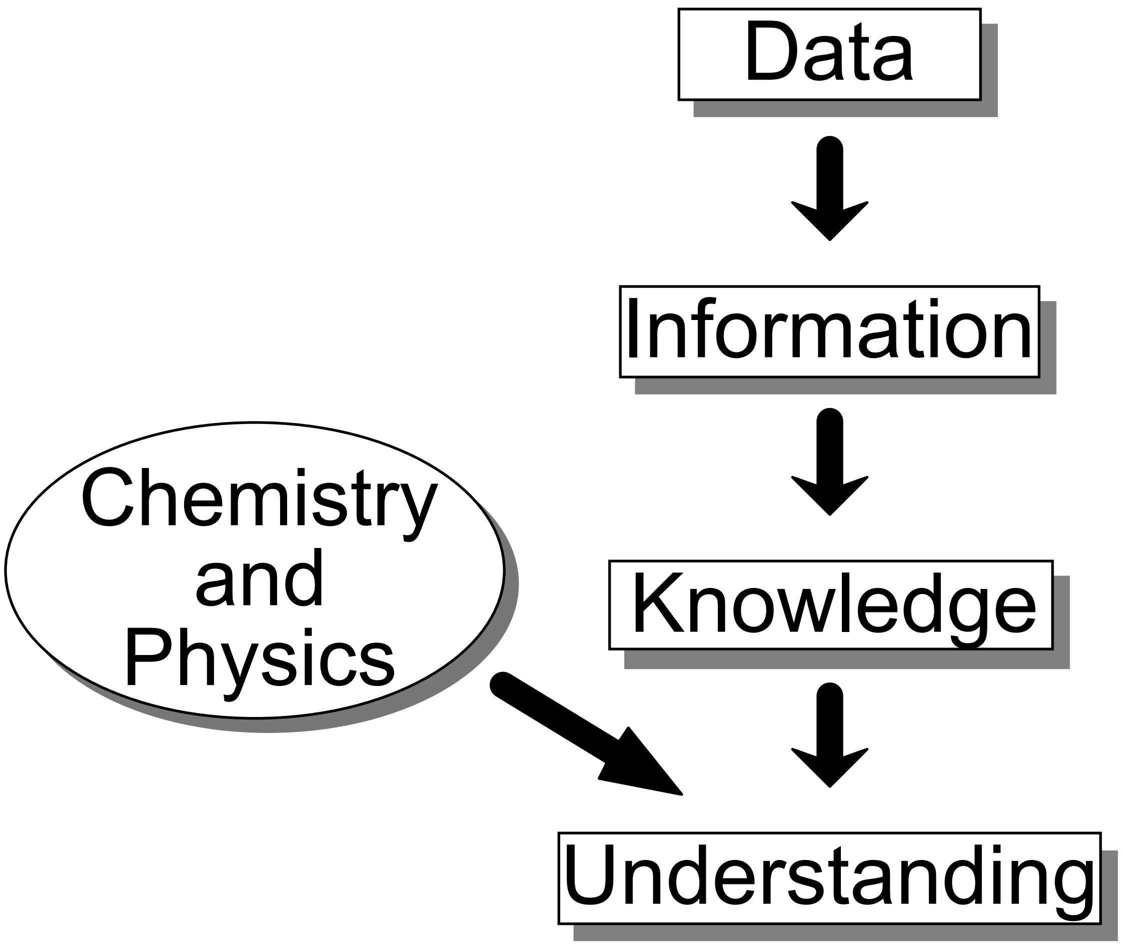
Multi-way Analysis

Multi-way Analysis is concerned with data that is measured as a function of *three or more factors*.

Multivariate Images

A data array of *dimension three* (or more) where the first two dimensions are *spatial* and the last dimension(s) is a function of another variable.

Information Hierarchy

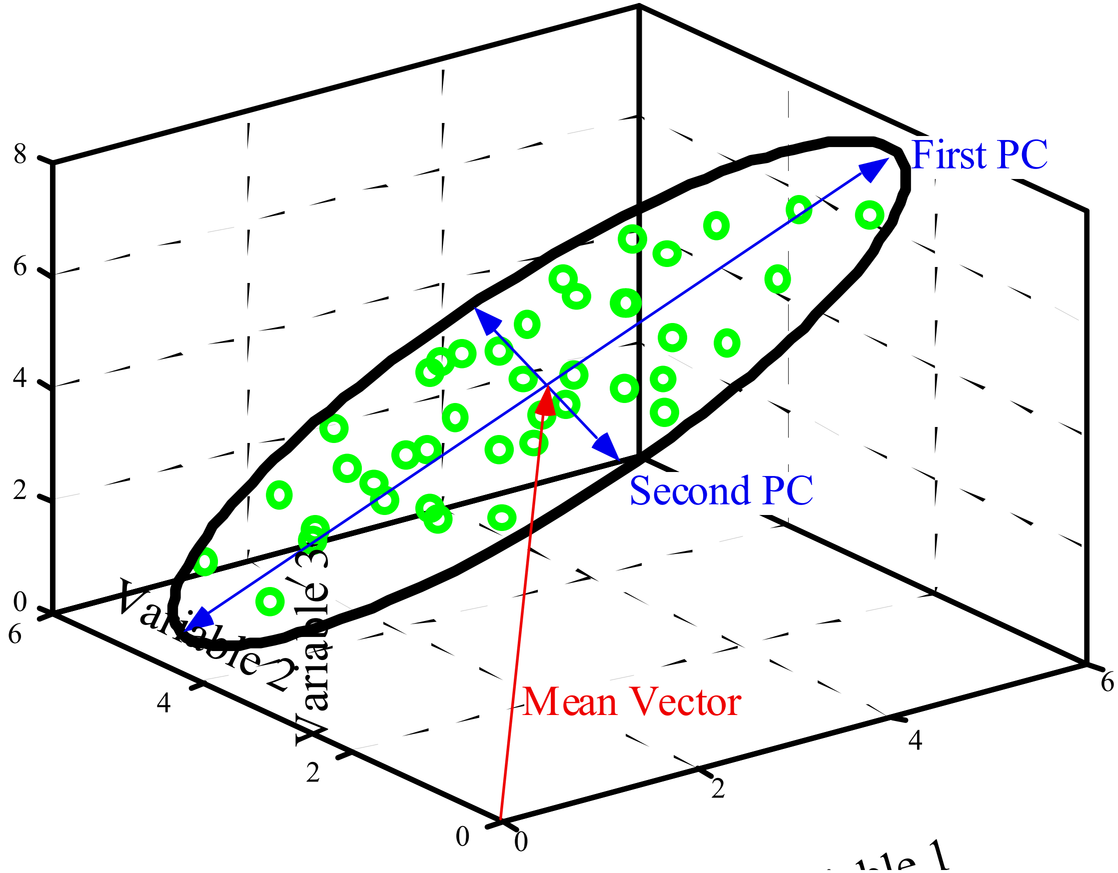


Why Chemometrics?

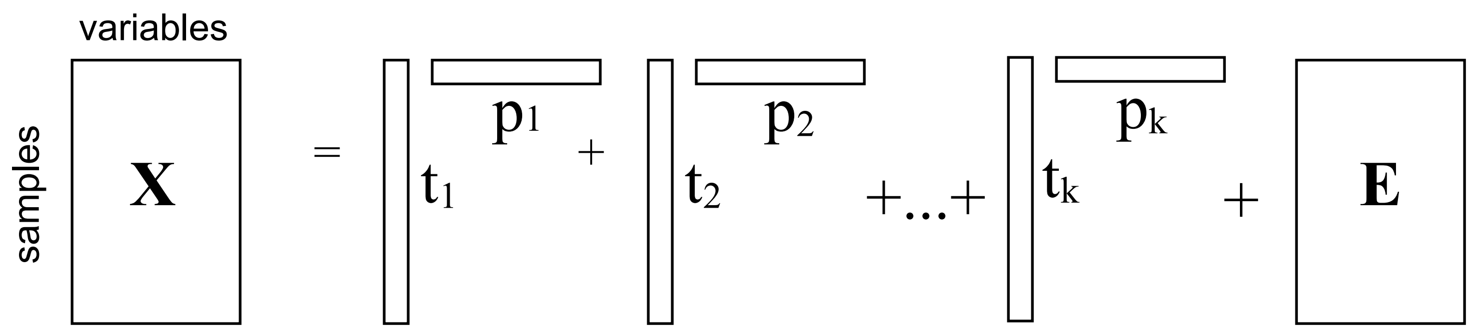
- It's a multivariate world!
  - Need windows into this multivariate world
- There are many things that simply can't be done if you don't recognize this, including
  - sample classification/pattern recognition
  - calibrations for complex systems (often spectroscopy)
  - transfer of calibrations between instruments
  - fault and upset detection
- Chemometrics focuses on the part of math and statistics applicable to *chemical* problems
- More expensive to do things with hardware if you can do them with math instead

Tools of the Trade

Principal Components Analysis



PCA Math



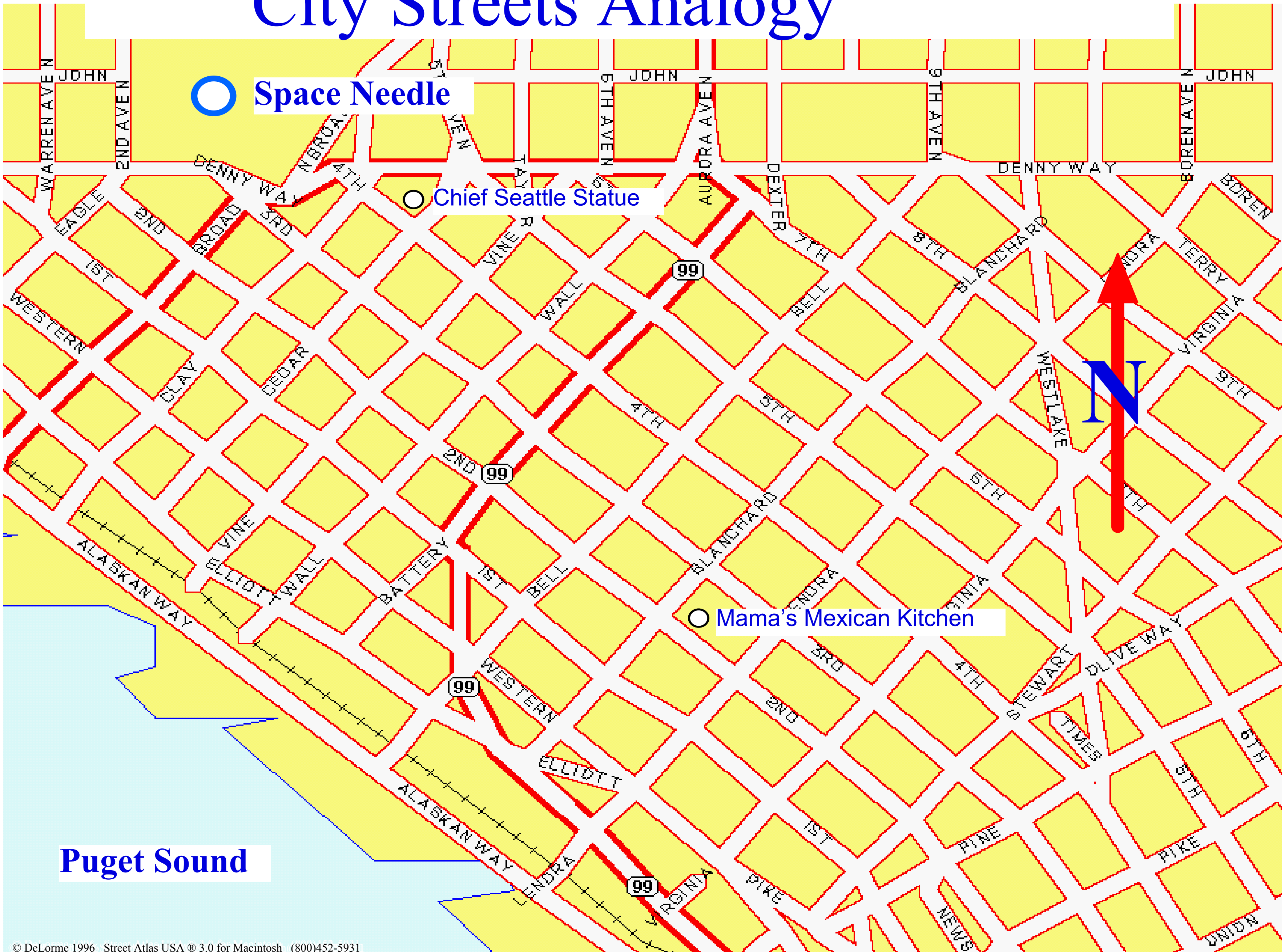
The $\mathbf{p}_i$ are the eigenvectors of the covariance matrix

$$\mathrm{cov}(\mathbf{X}) = \frac{\mathbf{X}^{\mathrm{T}}\mathbf{X}}{m-1}$$

$$\mathrm{cov}(\mathbf{X})p_i = \lambda_i p_i$$

and the λ_i are the eigenvalues. Amount of variance captured by $\mathbf{t}_i\mathbf{p}_i$ proportional to λ_i.

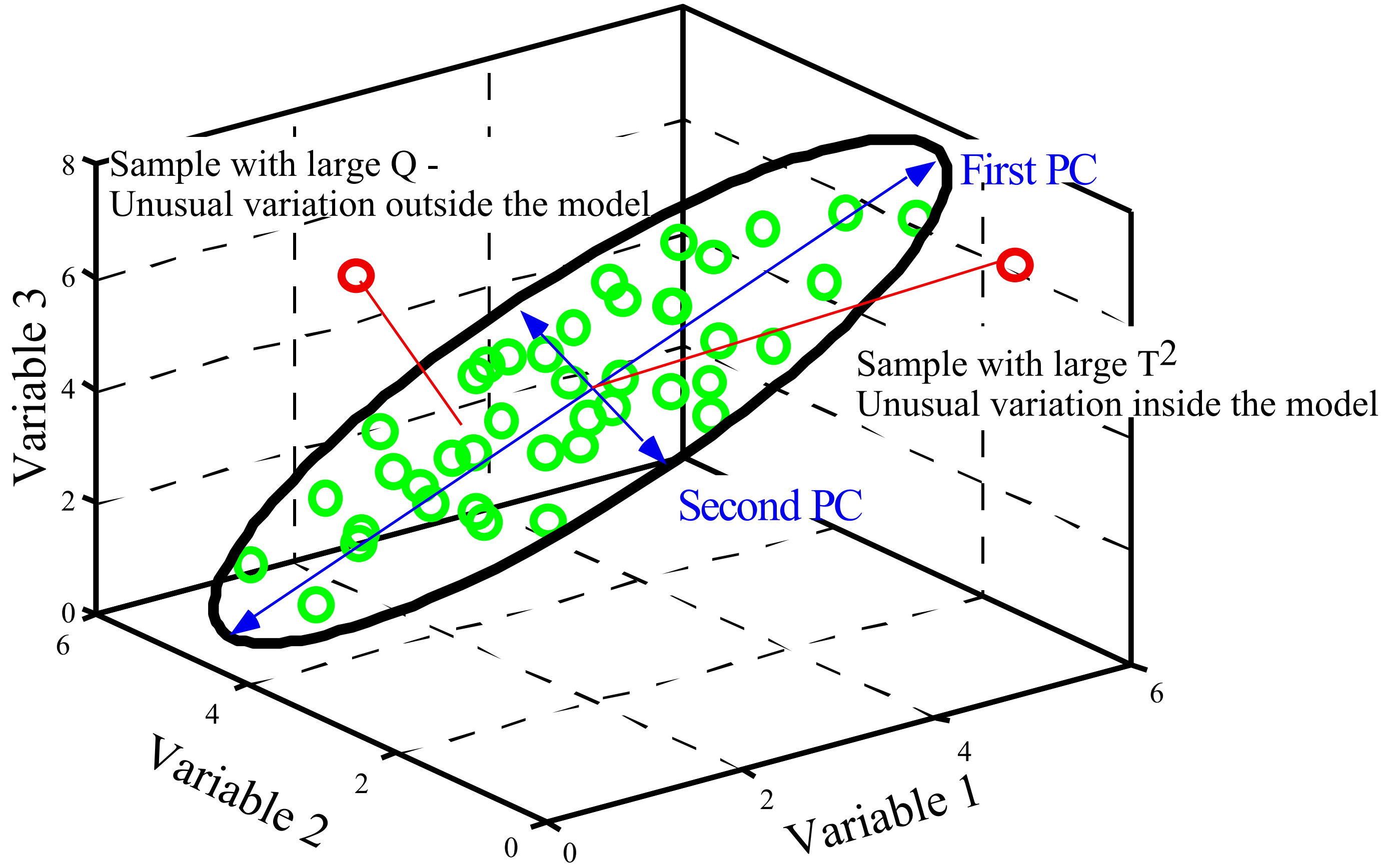
City Streets Analogy



Properties of PCA

- $\mathbf{t}_i,\mathbf{p}_i$ pairs ordered by amount of *variance captured*
- *variance = information*
- $\mathbf{t}_i$ or *scores* form an orthogonal set $\mathbf{T}_k$ which describe relationship between *samples*
- $\mathbf{p}_i$ or *loadings* form an orthonormal set $\mathbf{P}_k$ which describe relationship between *variables*

Geometry of Q and T^2



EIGENVECTOR RESEARCH INCORPORATED

PCA Statistics

Control limits can be developed for the lack of model fit statistic Q:

$$Q_i = \mathbf{e}_i \mathbf{e}_i^T = \mathbf{x}_i(\mathbf{I} - \mathbf{P}_k \mathbf{P}_k^T)\mathbf{x}_i^T$$

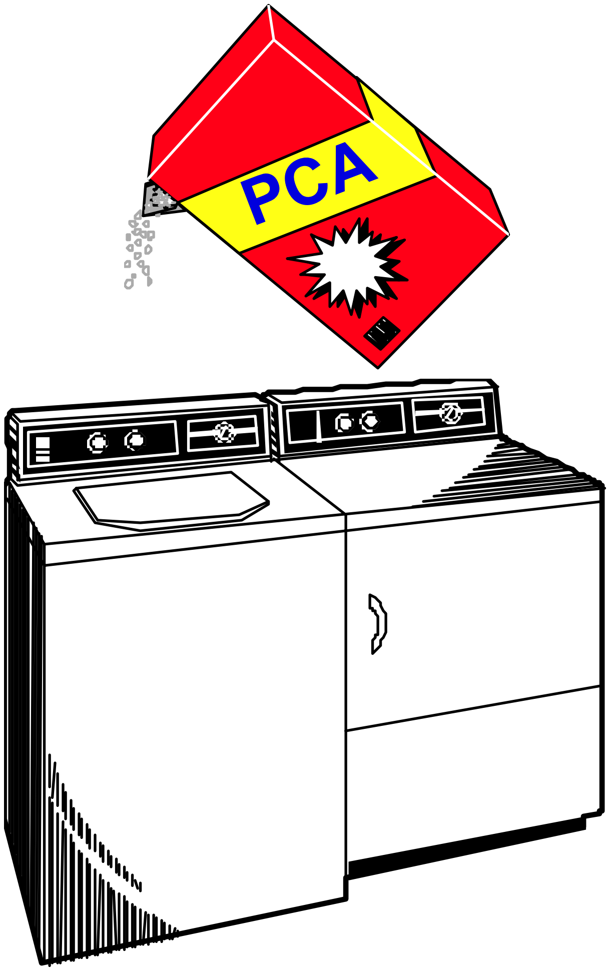
and Hotelling's T^2 statistic:

$$T_i^2 = \mathbf{t}_i \lambda^{-1} \mathbf{t}_i^T = \mathbf{x}_i \mathbf{P}_k \lambda^{-1} \mathbf{P}_k \mathbf{x}_i^T$$

Control limits can also be developed for the individual scores (t_{ij}) and the residuals (e_{ij})

Dirty T-shirt Analogy

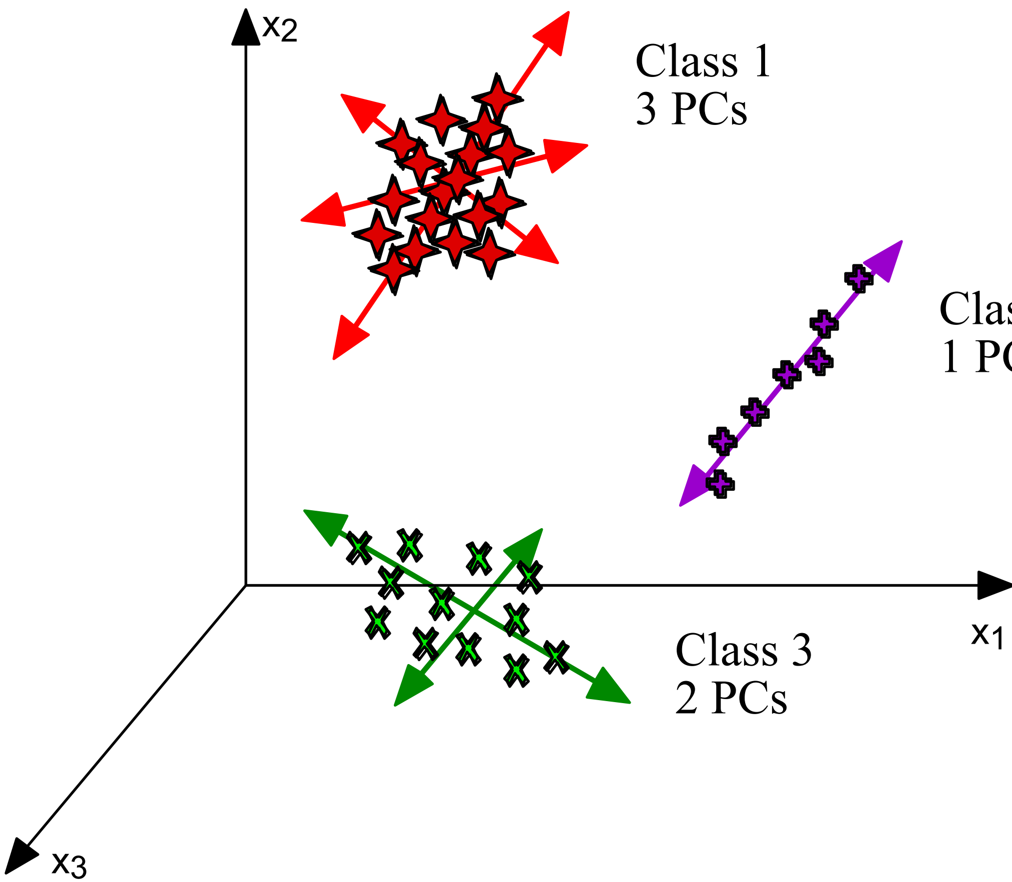
PCA attempts to partition data into deterministic and non-deterministic portions



Applying a PCA Model to New Data

- *A PCA model is a description of a data set*, including its mean, amount of variance and its direction, dimensionality, and typical residuals
- New data can be compared with existing PCA models to see if it is "similar"
- Used in Multivariate Statistical Process Control (MSPC)

SIMCA



Regression

- Often want to obtain a relationship between one set of variables, **X**, and another, **y** or **Y**.
  - Absorbances -> concentrations or other property
  - Acoustic signature -> particle size distribution
- Want **y** = **Xb** + **e** (or **Y** = **XB** + **E**)
- Relationship may be non-causal
- May have more variables than samples
- Highly collinear data
- Problem if using MLR!

Estimation of b: MLR

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

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

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

- There are many ways to obtain a pseudo-inverse, most obvious is Multiple Linear Regression (MLR), a.k.a. Ordinary Least Squares (OLS)
- In this case, $\mathbf{X}^{+}$ defined by:

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

EIGENVECTOR RESEARCH INCORPORATED

Problem with MLR

- Matrix inverse exists only if
  - Rank(X) = number of variables, but rank(X) ≤ min {mx,nx}
  - **X** has more samples than variables (problem with spectra)
  - Columns of **X** are not collinear
- Matrix inverse may exist but be highly unstable if **X** is nearly rank deficient
- Much of multivariate calibration involves tricks for obtaining regression models in spite of problems with matrix inverses!

Getting Around the MLR Problem

- MLR doesn't work when mx < nx, or when variables are colinear
- Possible solution: eliminate variables, *e.g.* stepwise regression or other variable selection
  - how to choose which variables to keep?
  - 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

- PCR is one way to deal with ill-conditioned regression problems.
- Property of interest **y** is regressed on PCA scores:

$$\mathbf{X}^{+} = \mathbf{P}_k(\mathbf{T}_k\mathbf{T}_k^{\mathrm{T}})^{-1}\mathbf{T}_k^{\mathrm{T}}$$

- Problem is to determine k, the number of PCs to retain in formation of $\mathbf{X}^{+}$

Determining the Number of Factors (PCs or LVs)

- A central idea in PCR (and PLS) is that variance is important: use factors that describe lots of variance first
- Question: when do you stop?
- Answer: use *cross-validation*
- Build model on part of the data and use remaining data to test model as a function of number of factors retained

Model Cross-validation and Validation

- Cross-validation is a common step in model building
- Models should also be validated on totally separate data sets if possible
- Why is this important?
- *It is very easy to fit data, but making predictions is hard!*

Problem with PCR

- Some PCs not relevant for prediction, only relevant for describing **X**
- Result of determining PCs without regard to property to be predicted
- Solution: find factors using some information from **y** (or **Y**), not just **X**

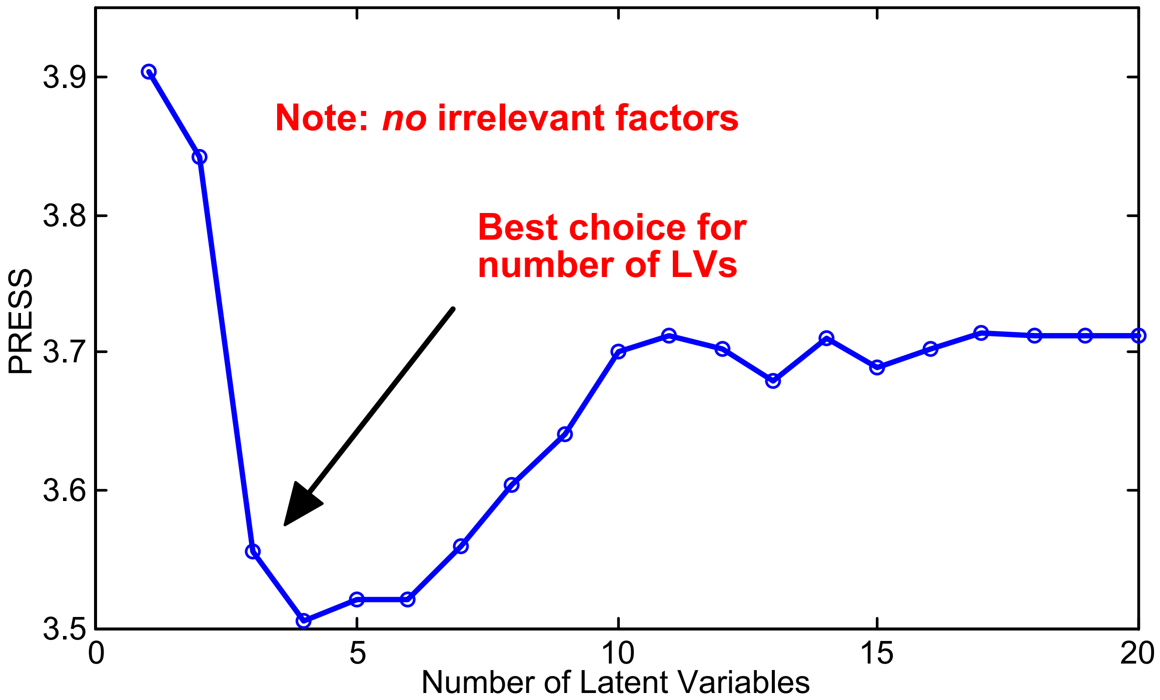
Solution: Partial Least Squares Regression (PLS)

- PLS is related to PCR and MLR
  - PCR captures maximum variance **X**
  - MLR achieves maximum correlation with **y**
  - PLS tries to do both, maximizes covariance
- PLS requires addition of weights **W** to maintain orthogonal scores
- Factors calculated sequentially by projecting **y** through **X**
- Matrix inverse is:

$$\mathbf{X}^{+} = \mathbf{W}_k(\mathbf{P}_k^{T}\mathbf{W}_k)^{-1}(\mathbf{T}_k^{T}\mathbf{T}_k)\text{-}1\,\mathbf{T}_k^{T}$$

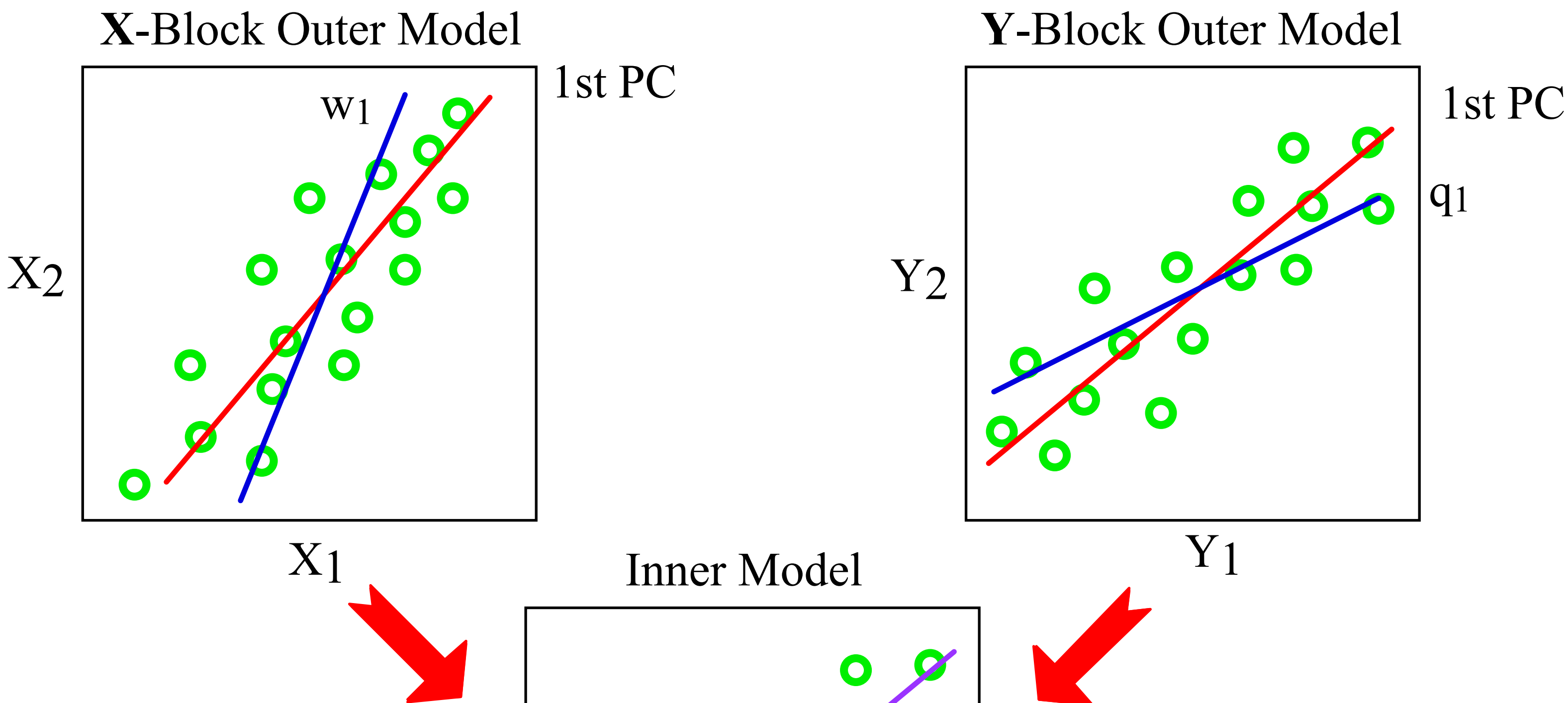
EIGENVECTOR RESEARCH INCORPORATED

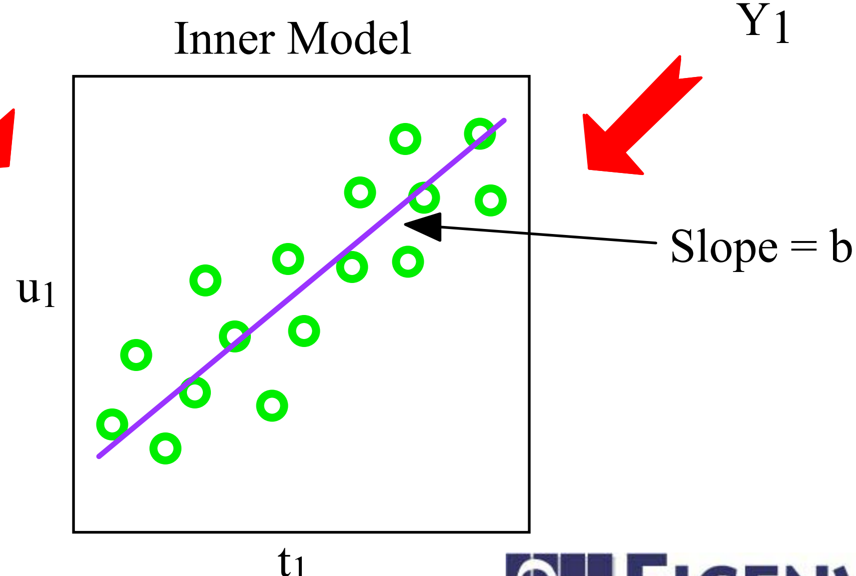
Cross-validation PRESS Curve



EIGENVECTOR RESEARCH INCORPORATED

PLS2 Modelling





$\mathbf{w}_1$ and $\mathbf{q}_1$ are similar to first PCs in **X** and **Y** but are rotated so that there is better correlation between the **X** scores $\mathbf{t}_1$ (= $\mathbf{X}\mathbf{w}_1$) and **Y** scores $\mathbf{u}_1$ (= $\mathbf{Y}\mathbf{q}_1$)

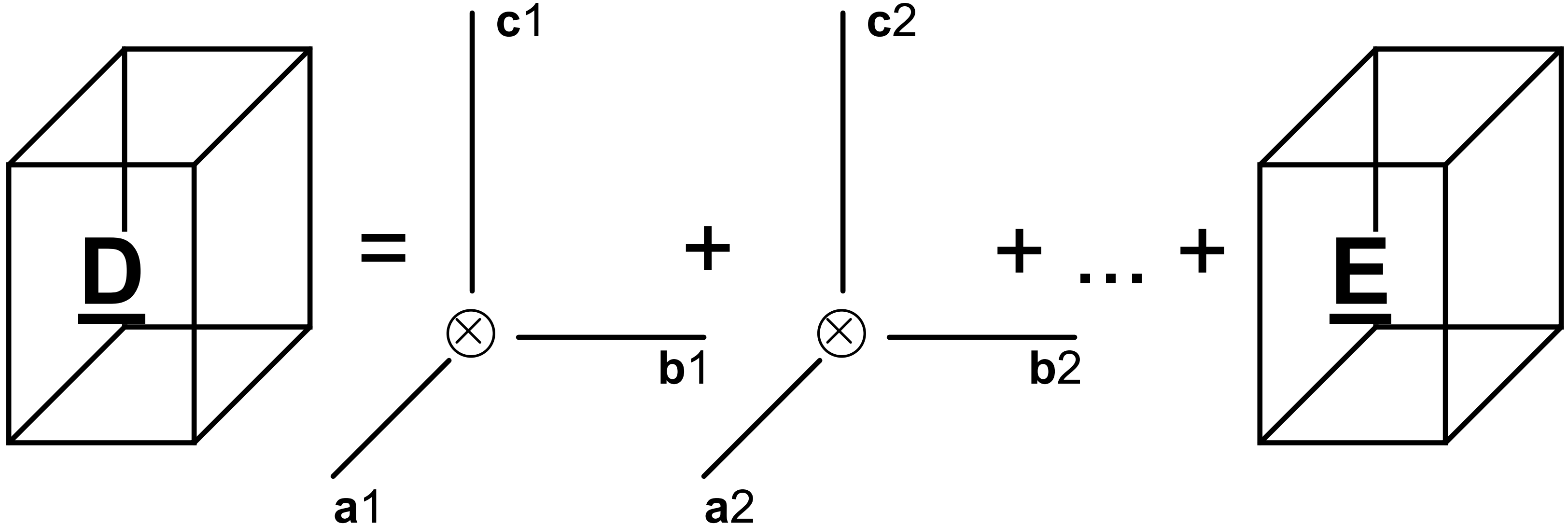
Multivariate Curve Resolution

- MCR attempts to extract pure component spectra and concentration profiles evolving systems like GC-MS
- Given a response matrix $\mathbf{N}_m$ that is the product of concentration profiles **C** and pure component spectra **S**:

$$\mathbf{N}_m = \mathbf{CS} + \mathbf{E}$$

- Uses alternating and constrained least squares to get **C** and **S**

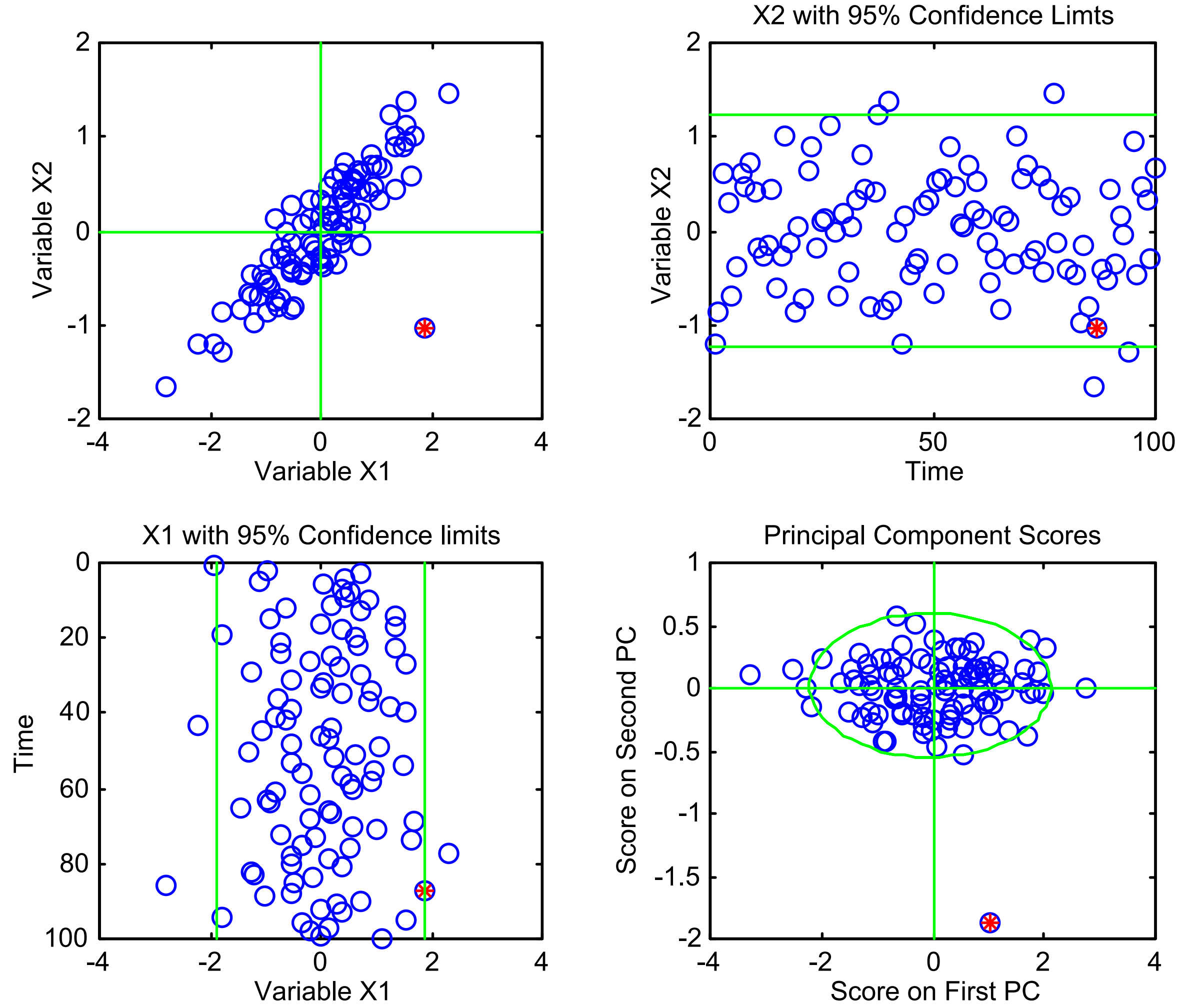
The PARAFAC Model



EIGENVECTOR RESEARCH INCORPORATED

Opportunities in Process Analytical Technology (PAT)

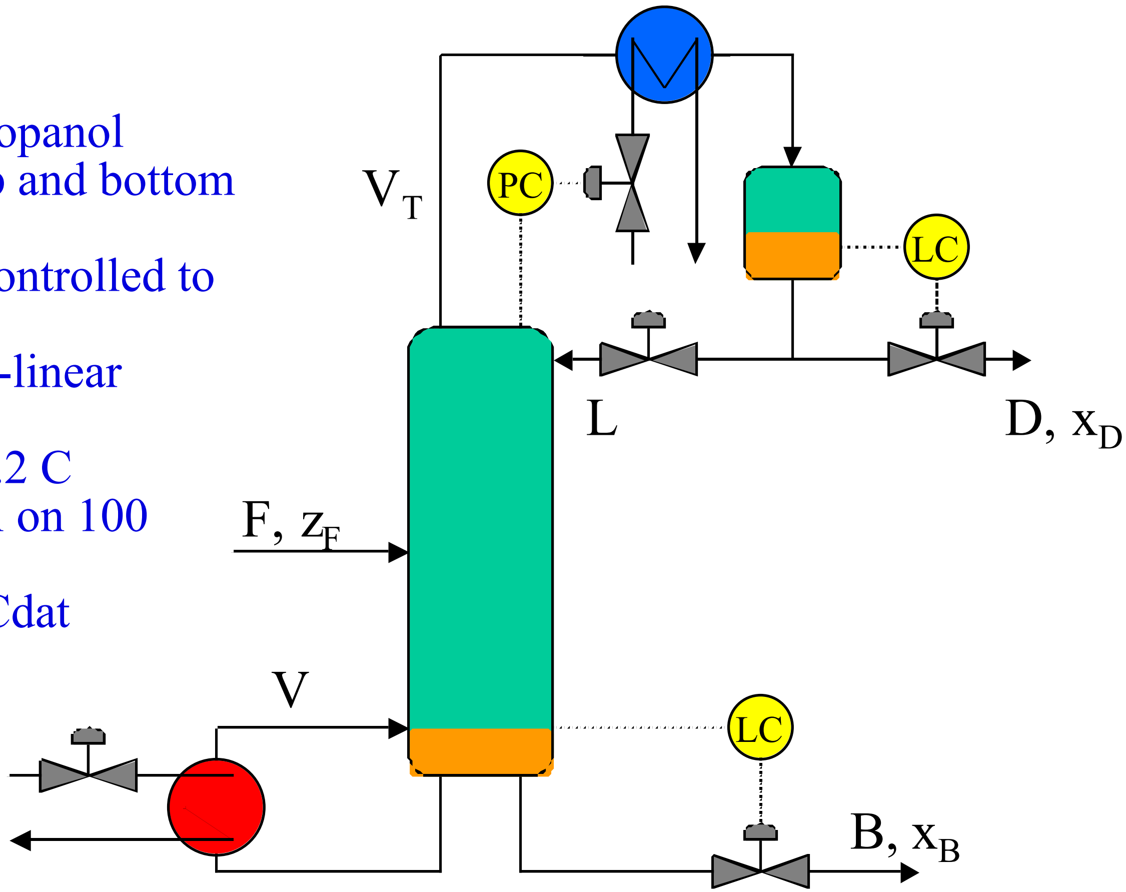
Multivariate Statistical Process Control



EIGENVECTOR RESEARCH INCORPORATED

Example from Distillation

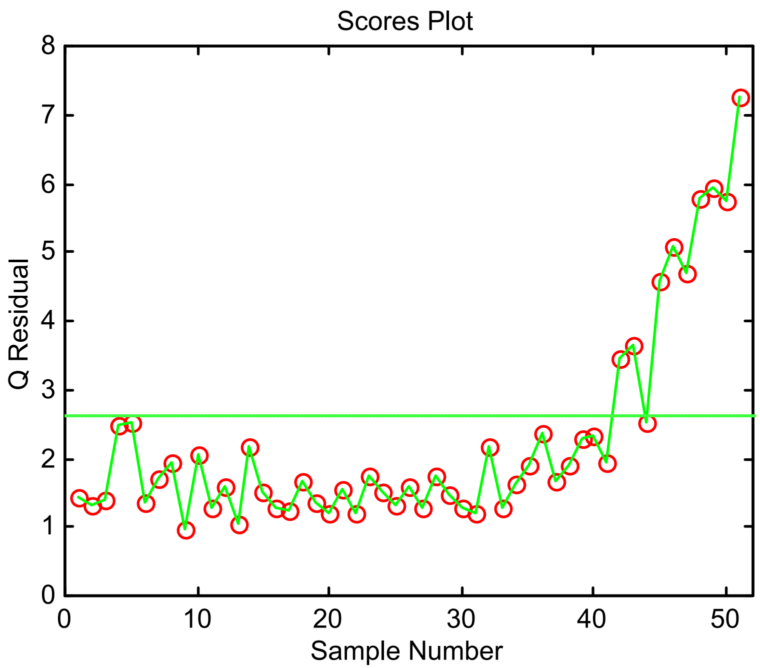
- 41 stage column
- hexane and isopropanol
- LV control of top and bottom compositions
- top and bottom controlled to 99% purity
- full dynamic non-linear simulation
- noise on temps 0.2 C
- build PCA model on 100 normal samples
- load columMSPCdat

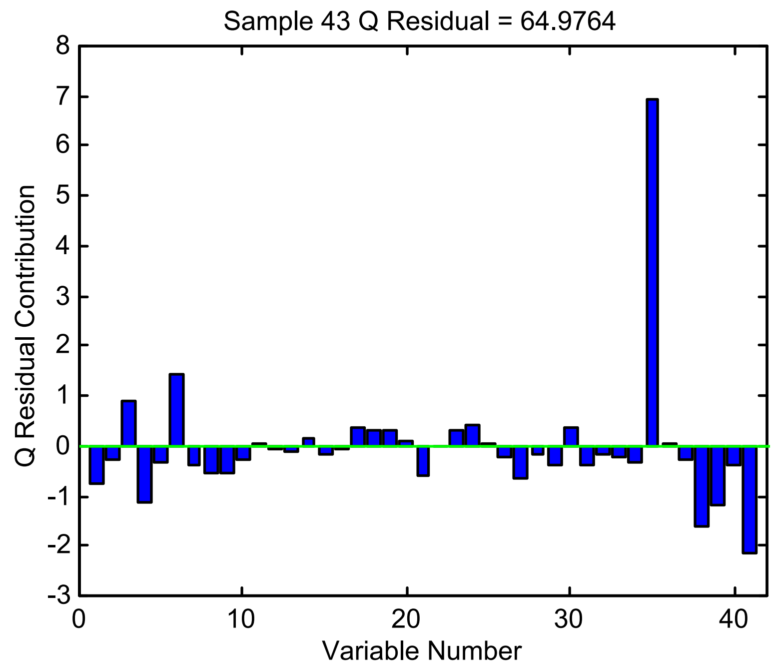


EIGENVECTOR RESEARCH INCORPORATED

Fault #1: Temperature Sensor

- Ramped bias (0.2 to 2 C) is added to temperature from tray 35 at sample 31

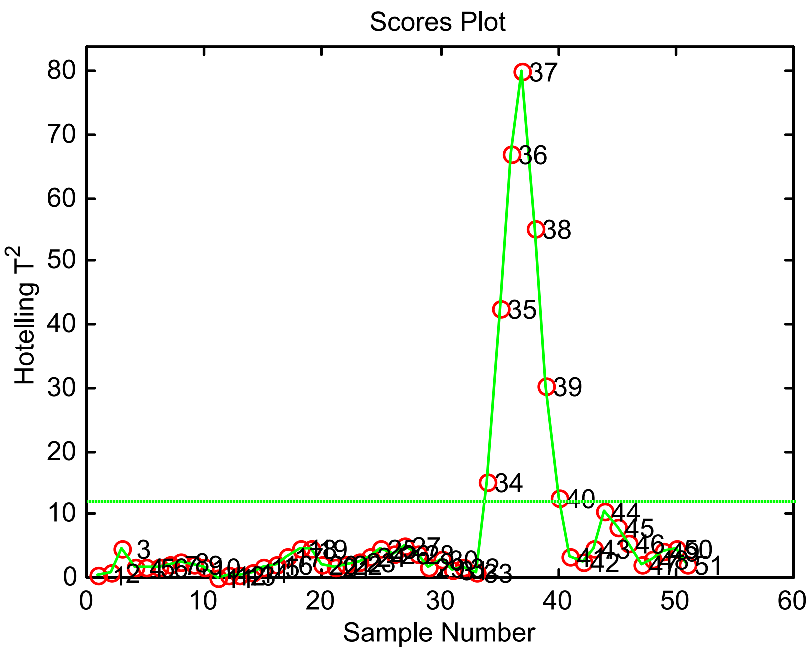


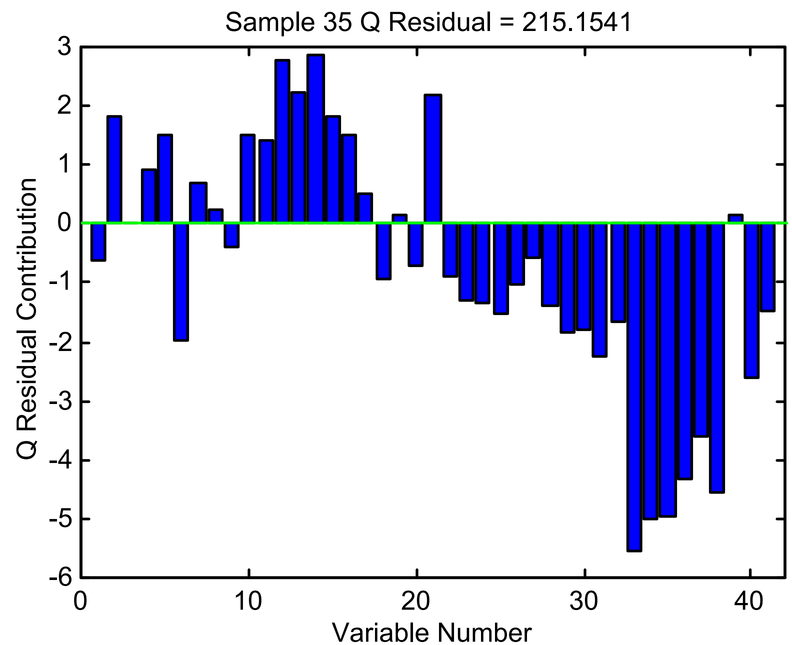


EIGENVECTOR RESEARCH INCORPORATED

Fault #2: Feed Quality

- Amount of feed entering as vapor goes from 0% to 50% at time 31





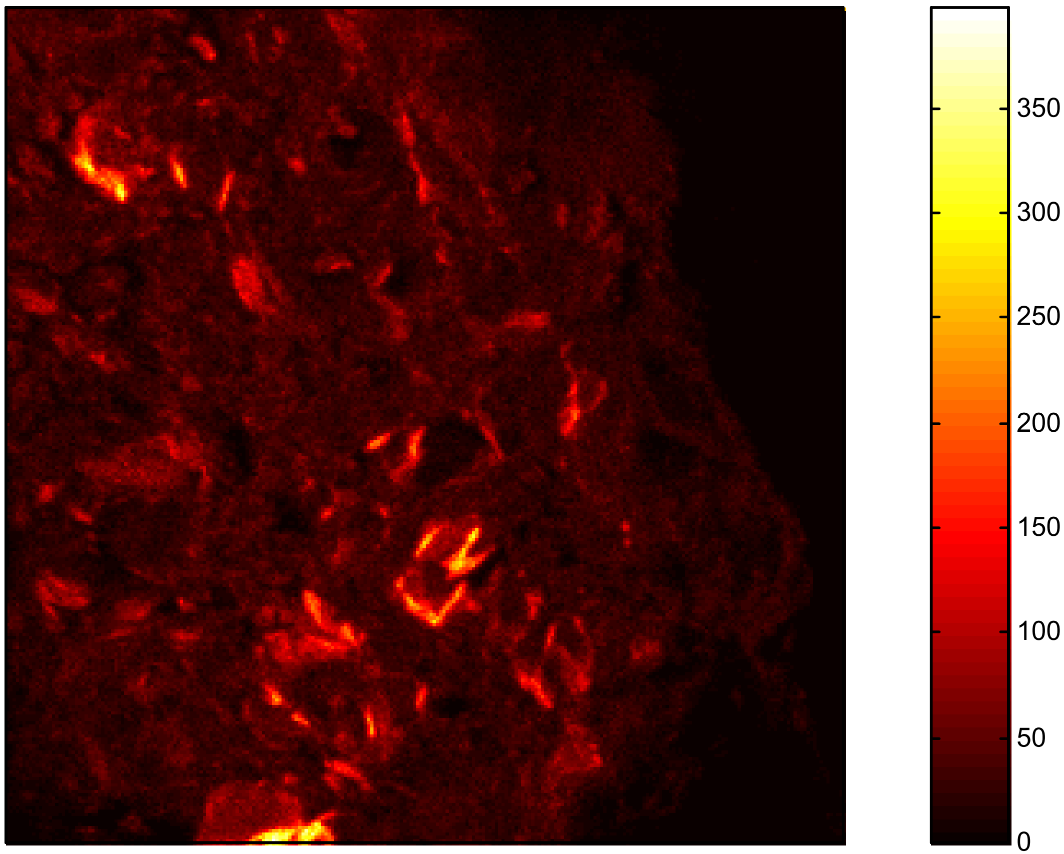
EIGENVECTOR RESEARCH INCORPORATED

TOF-SIMS of Time Release Drug Delivery System

- Multilayer drug beads serve as controlled-release delivery system
- TOF-SIMS taken of cross section of bead
- Evaluate integrity of layers, distribution of ingredients
- Thanks again to Physical Electronics and Anna Belu for the data!

Reference: A.M. Belu, M.C. Davies, J.M. Newton and N. Patel, "TOF-SIMS Characterization and Imaging of Controlled-Release Drug Delivery Systems, Anal. Chem., 72(22), pps 5625-5638, 2000

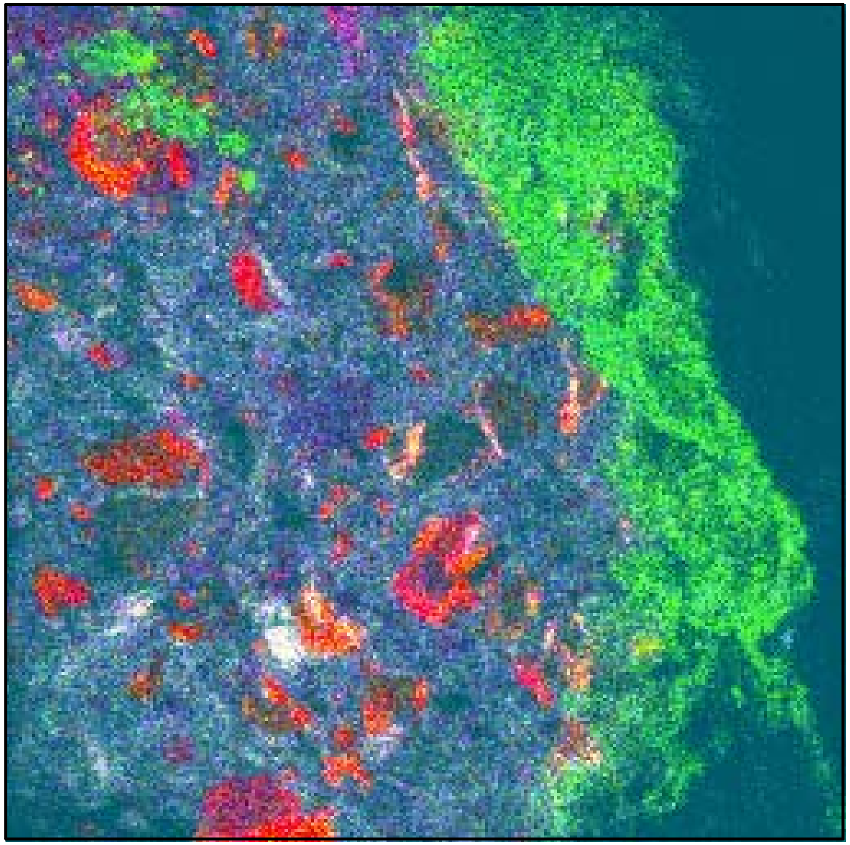
Total Ion Image of Bead



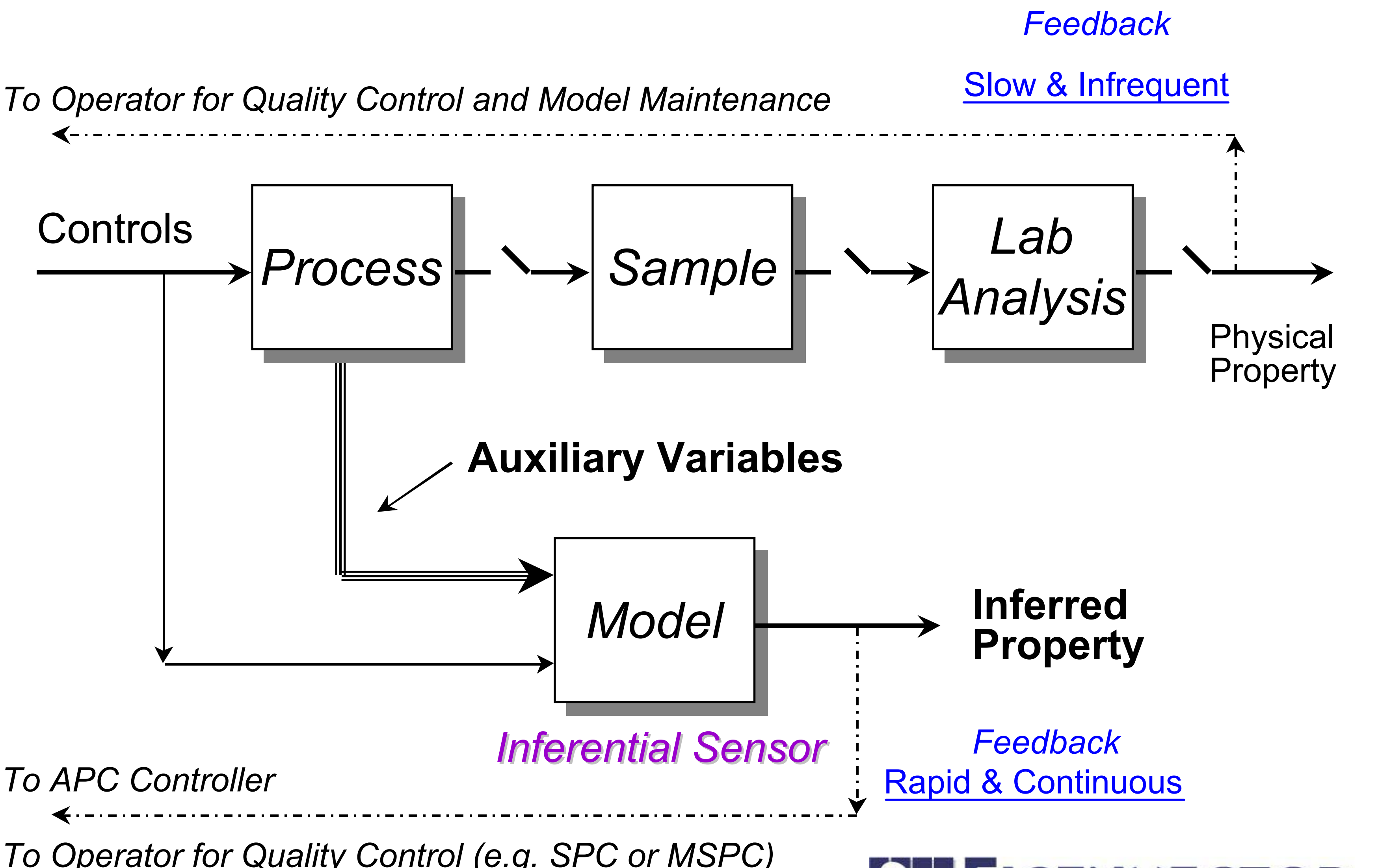
EIGENVECTOR RESEARCH INCORPORATED

False Color Image based on Scores of First 3 PCs

False Color Image of First 3



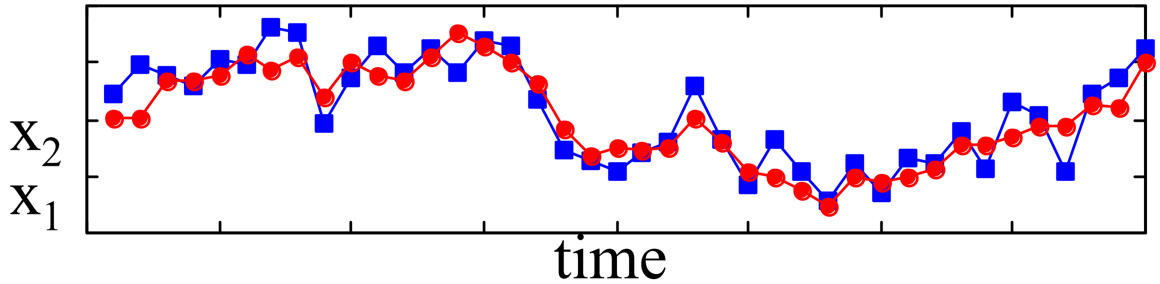
Inferential Measurements



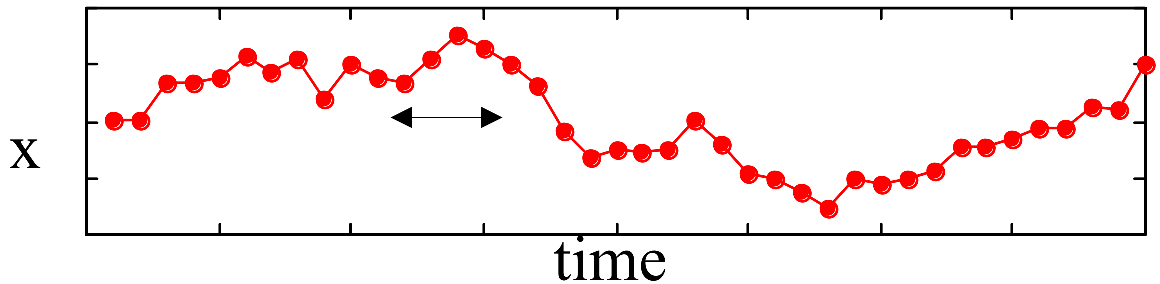
EIGENVECTOR RESEARCH INCORPORATED

Process Data Characteristics

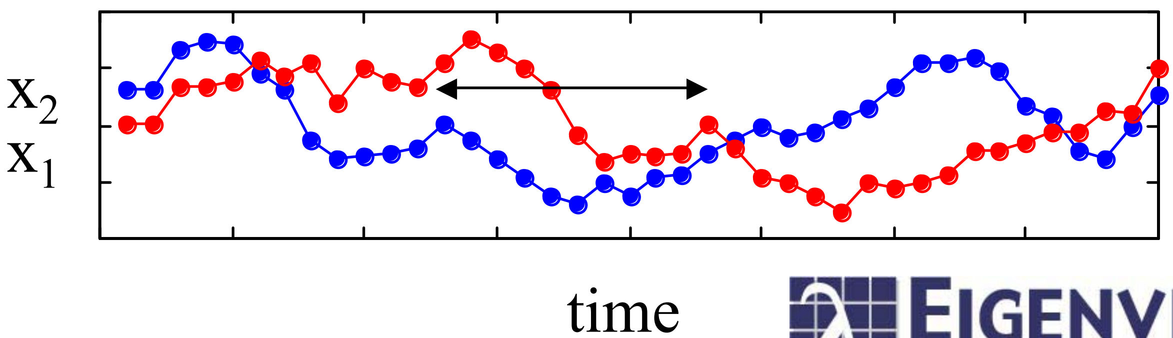
correlated: variables are not independent



autocorrelated: variables correlate with themselves over time



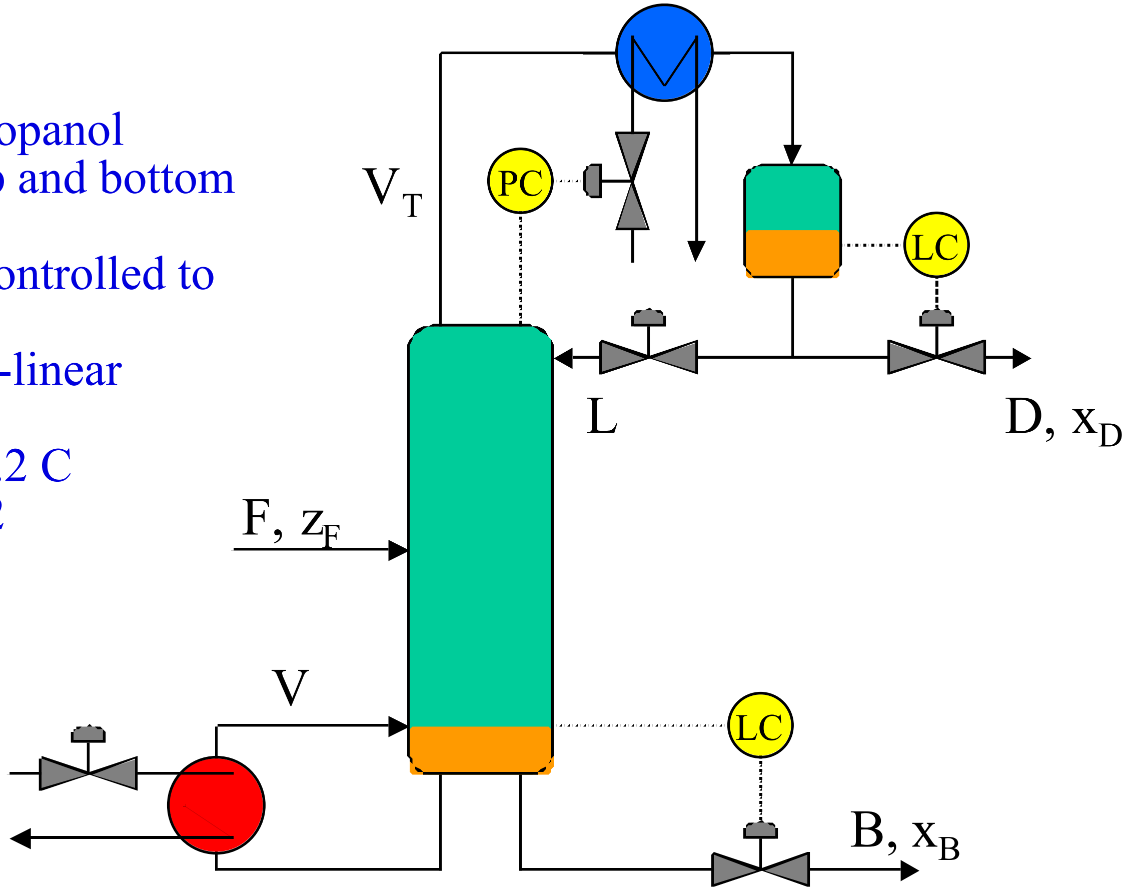
crosscorrelated: variables correlate with other variables at different time lags



EIGENVECTOR RESEARCH INCORPORATED

Distillation Column

- 41 stage column
- hexane and isopropanol
- LV control of top and bottom compositions
- top and bottom controlled to 99% purity
- full dynamic non-linear simulation
- noise on temps 0.2 C
- load DIST_EX_2

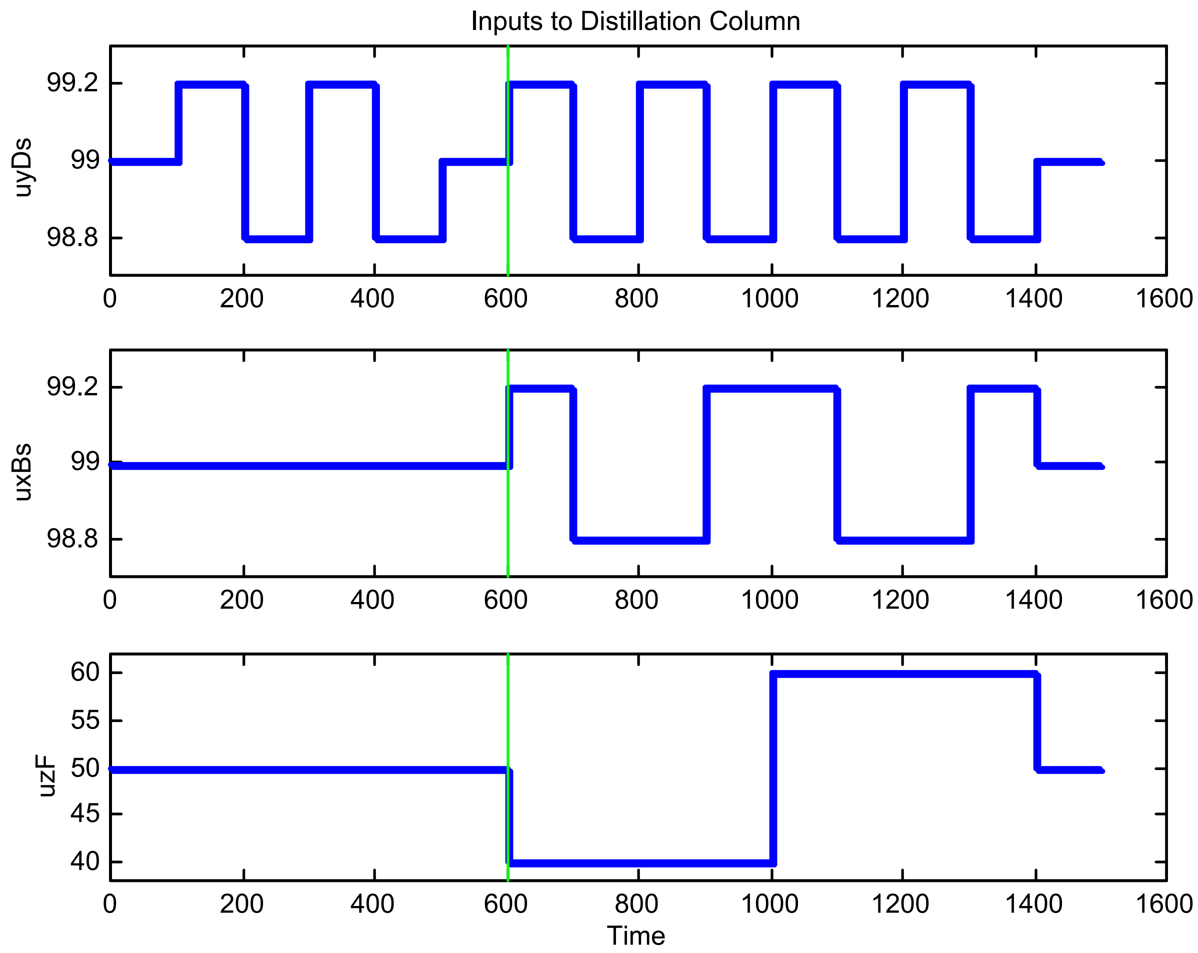


EIGENVECTOR RESEARCH INCORPORATED

Goal

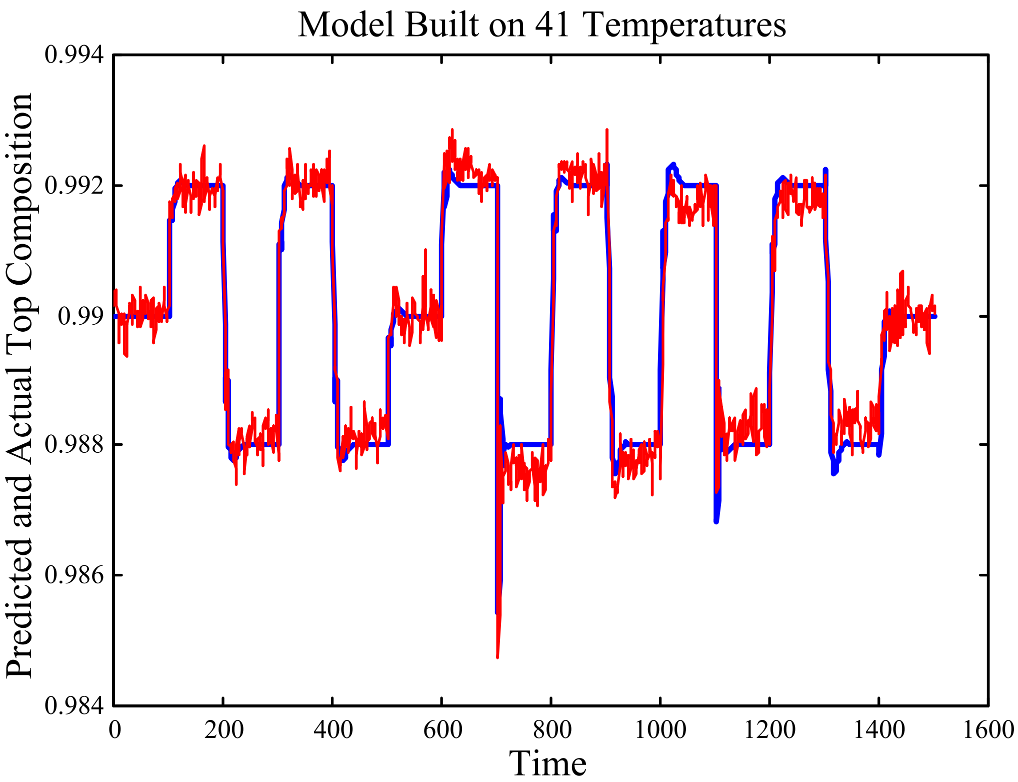
- Develop inferential sensor to predict distillate composition based on tray temperatures
- Make model work over a range of operating conditions
- Used designed experiment to generate data for identification of model
- Can use model for control and/or monitoring purposes

Designed Experiment



EIGENVECTOR RESEARCH INCORPORATED

If Disturbances are Included in Modeling Data, Model Works

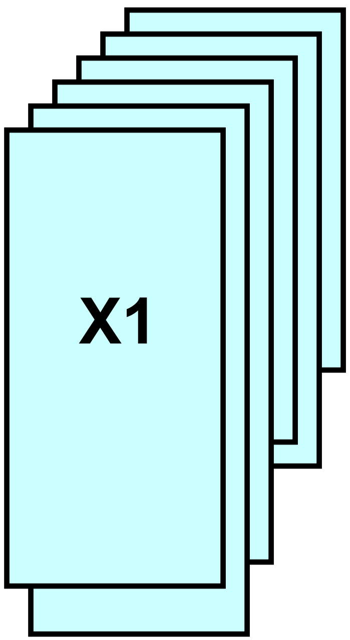


EIGENVECTOR RESEARCH INCORPORATED

Batch MSPC

- Multi-way methods can be used to monitor batches
- Build PARAFAC or PARAFAC2 model on normal data, apply to new batches
- Example from semiconductor etch process
- Problem: batches often of unequal length!

PARAFAC2 Model



The direct fitted PARAFAC2 model is:

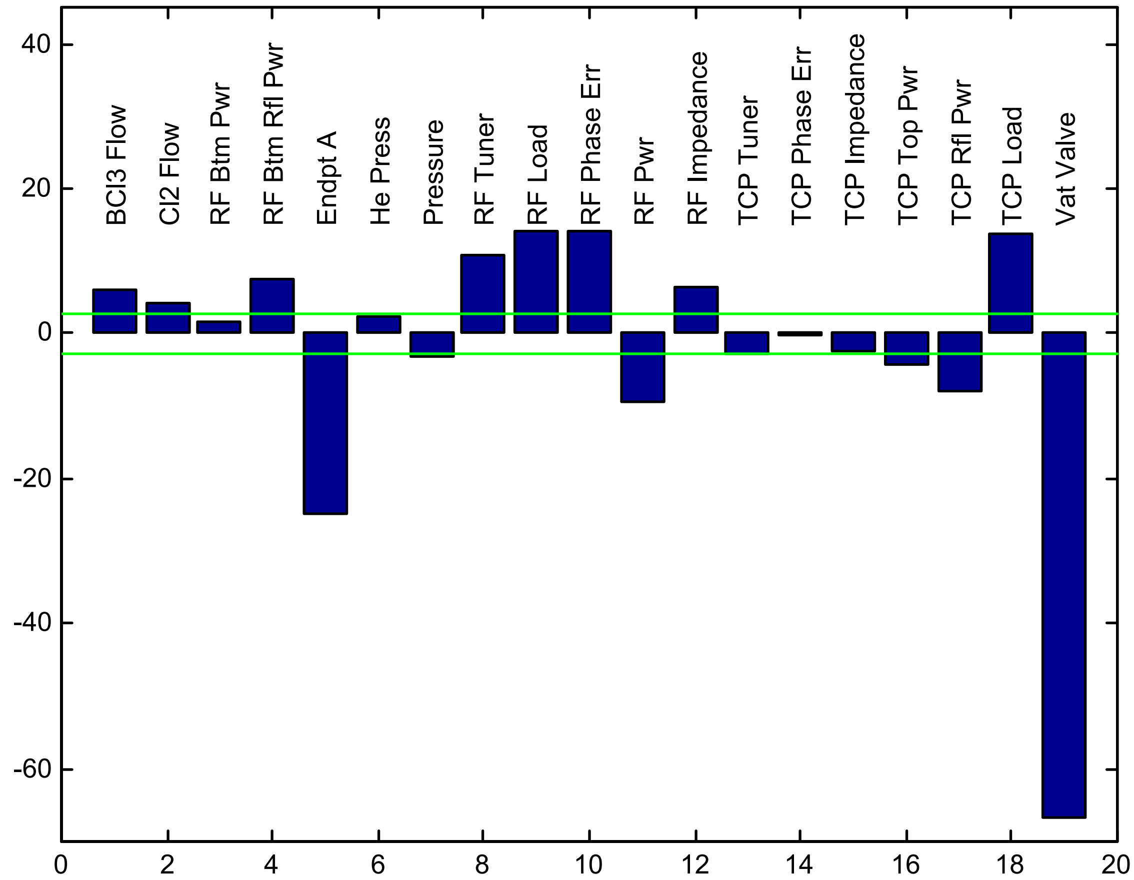
$$\mathbf{X}_k = \mathbf{F}_k\mathbf{D}_k\mathbf{A}^T + \mathbf{E}$$

subject to constraint that all $\mathbf{F}_k^T\mathbf{F}_k$ are equal. This is equivalent to the model

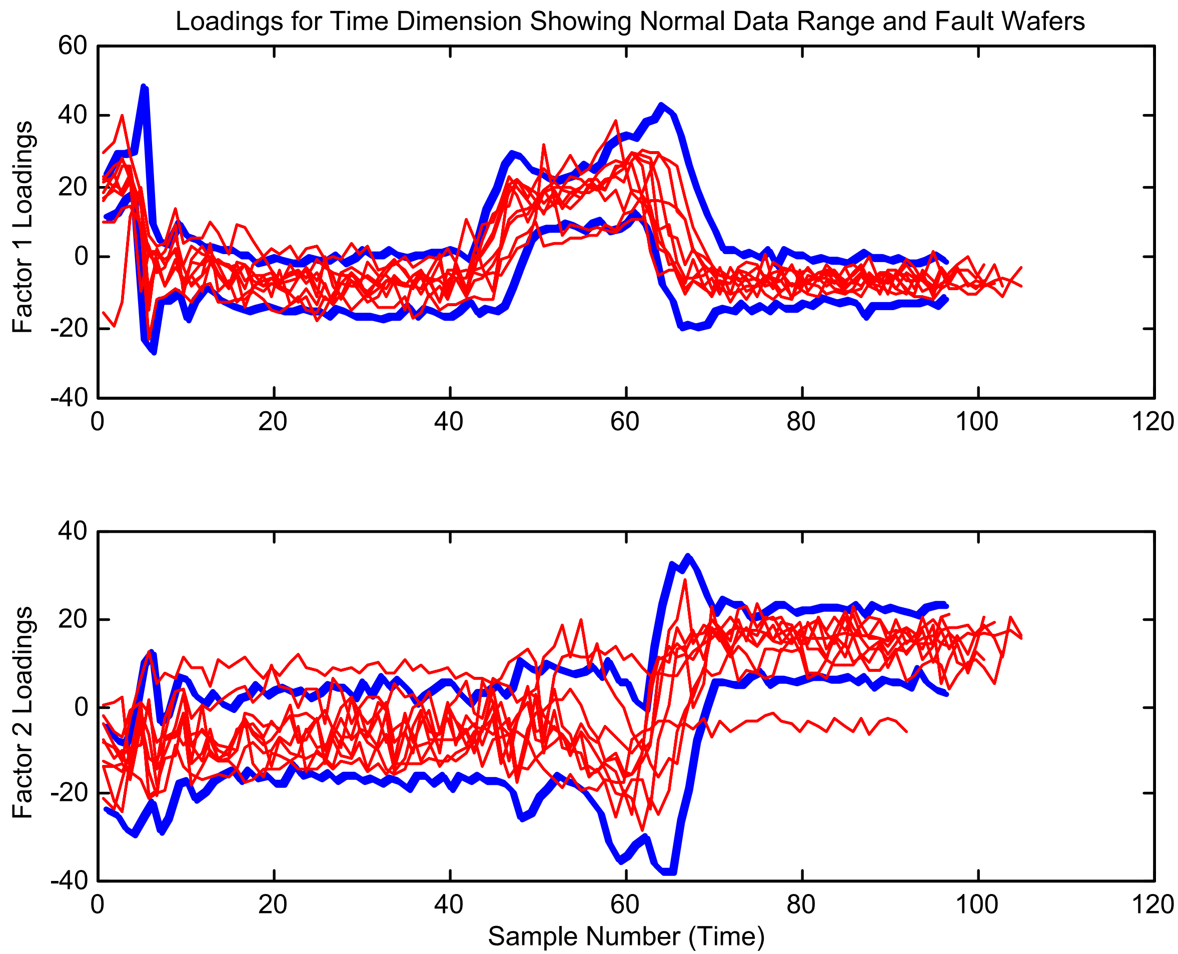
$$\mathbf{X}_k = \mathbf{P}_k\mathbf{F}\mathbf{D}_k\mathbf{A}^T + \mathbf{E}$$

where the $\mathbf{P}_k$ are orthonormal

PARAFAC2 Contributions



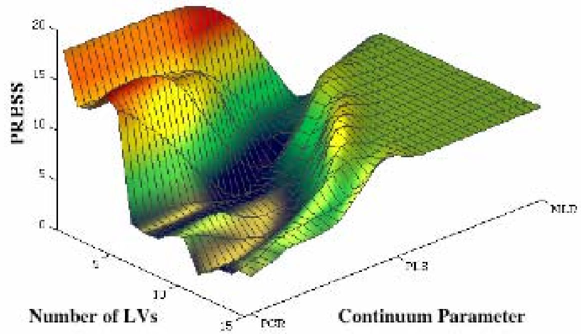
PARAFAC2 Loadings in Time Mode on New Batches



EIGENVECTOR RESEARCH INCORPORATED

Summary

- Chemometric tools emphasize
  - Interpretability
  - Predictive power
- Many places to use these tools in PAT
  - MSPC, BSPC
  - Calibrations, inferentials
  - Analysis of products



PLS_Toolbox 3.0

for use with MATLAB™

Barry M. Wise
Neal B. Gallagher
Rasmus Bro
Jeremy M. Shaver

EIGENVECTOR RESEARCH INCORPORATED

Contact Information

Eigenvector Research, Inc.
830 Wapato Lake Road
Manson, WA 98831
Phone: (509)687-2022
Fax: (509)687-7033
Email: bmw@eigenvector.com
Web: eigenvector.com

EIGENVECTOR RESEARCH INCORPORATED