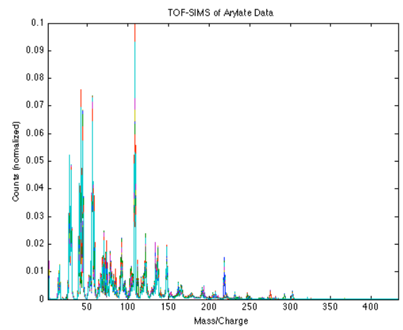


Multivariate Analysis for TOF-SIMS

Barry M. Wise

Contact Information

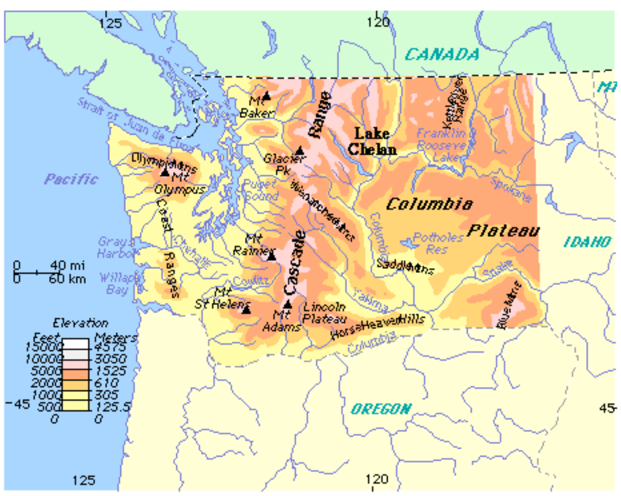
Eigenvector Research, Inc.
3905 West Eaglerock Drive
Wenatchee, WA 98801 USA
web: www.eigenvector.com

Barry M. Wise, Ph.D.
President
e-mail: bmw@eigenvector.com
phone: 509-662-9213

United States

EIGENVECTOR RESEARCH INCORPORATED 3

Washington State



EIGENVECTOR RESEARCH INCORPORATED 4

Outline

- Thinking Multivariate
- General Principles
- Data Sets
- Pattern Recognition with Principal Components Analysis
- Preprocessing
- Supervised Pattern Recognition: Classification
- Analysis of Multivariate Images
- Self Modeling Mixture Analysis, aka Curve Resolution
- Clustering
- Conclusions

EIGENVECTOR RESEARCH INCORPORATED 5

Definition of 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
2) design or select optimal *measurement* procedures and experiments.

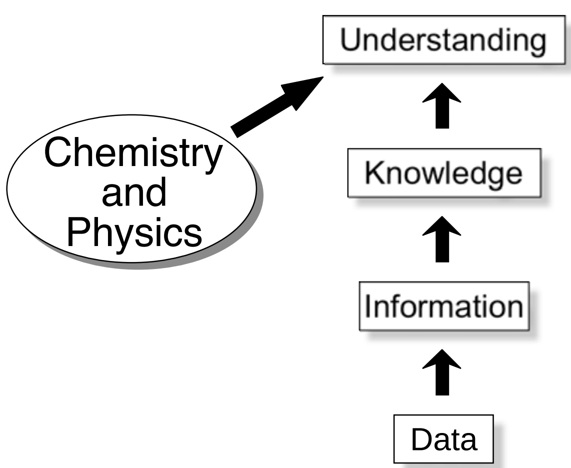
EIGENVECTOR RESEARCH INCORPORATED 6

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.

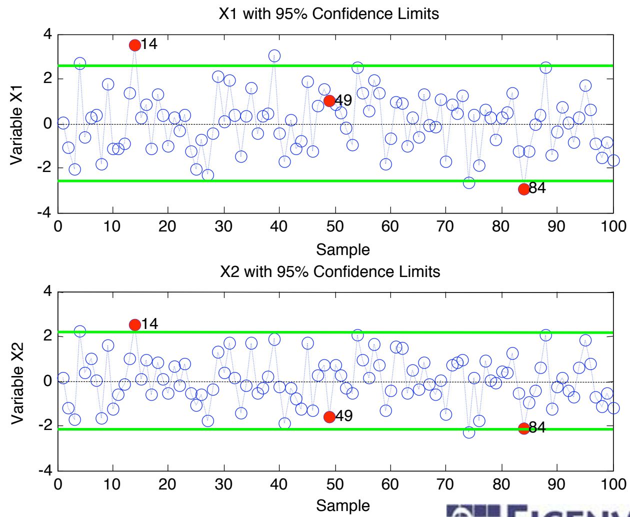
EIGENVECTOR RESEARCH INCORPORATED 7

Information Hierarchy



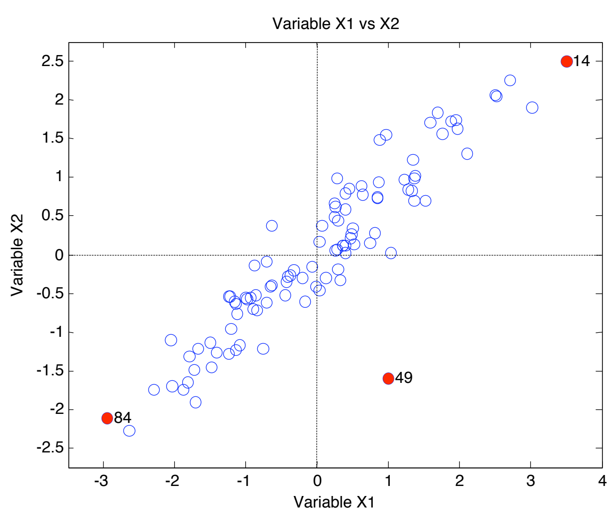
EIGENVECTOR RESEARCH INCORPORATED 8

Motivation: Which Point is Most Unique?



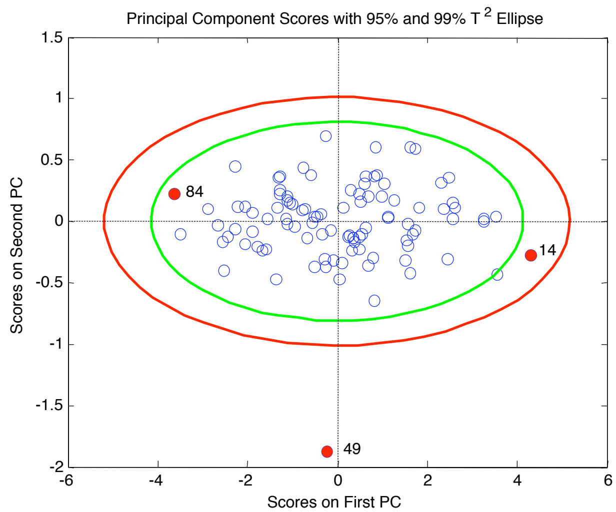
9

Plot X2 versus X1



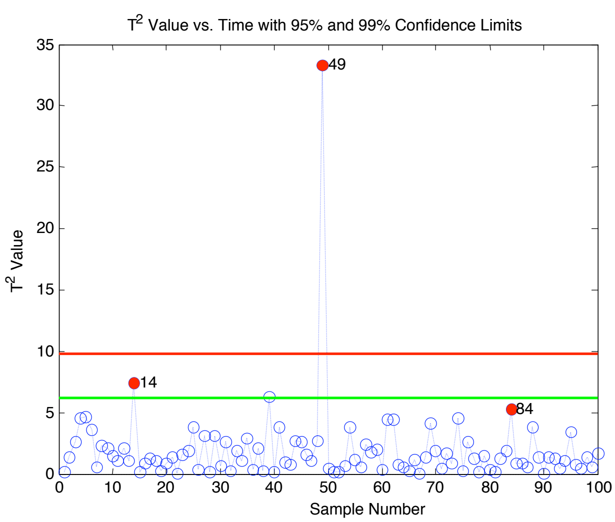
10

Principal Component Scores



EIGENVECTOR RESEARCH INCORPORATED 11

Monitor Single T^2 Chart



EIGENVECTOR RESEARCH INCORPORATED 12

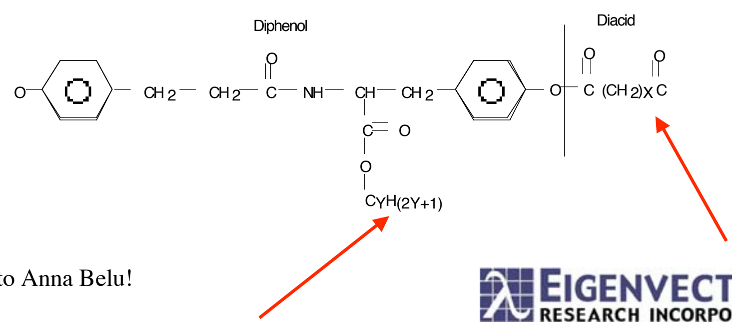
General Principles

- Balance
  - "Let the data speak for itself" - Bruce Kowalski
  - "Don't estimate what you already know" - John MacGregor
- Easier to fit data than predict it
  - Remember the parsimony principle
  - Validate models on independent test sets
- What you do before PCA, PLS etc. is critical
  - Experimental design, sample pedigree
  - Preprocessing to eliminate unwanted variance

EIGENVECTOR RESEARCH INCORPORATED 13

Example Data Set 1

- Tyrosine-derived polyarylates
  - From polymerization of diacids and diphenols
  - Backbone length varied (X)
  - Pendent (side) chain length varied (Y)

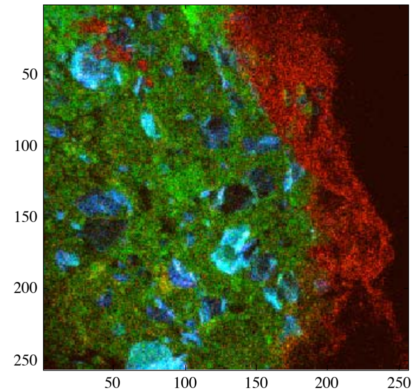


Thanks to Anna Belu!

EIGENVECTOR RESEARCH INCORPORATED 14

Example Data Set 2

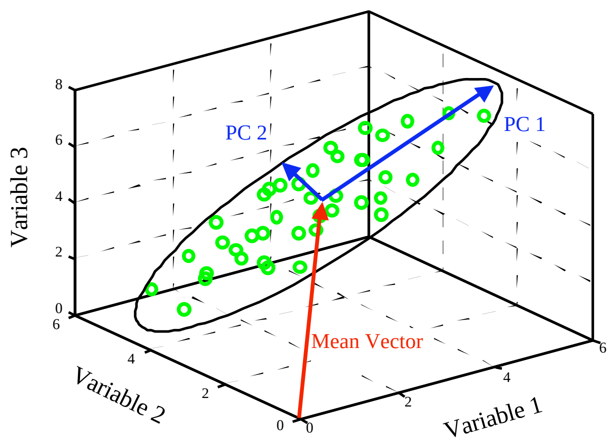
- Multilayer drug bead-controlled release delivery system
- TOF-SIMS taken of cross section of bead
- Evaluate integrity of layers, distribution of consituents



Thanks to Anna Belu!

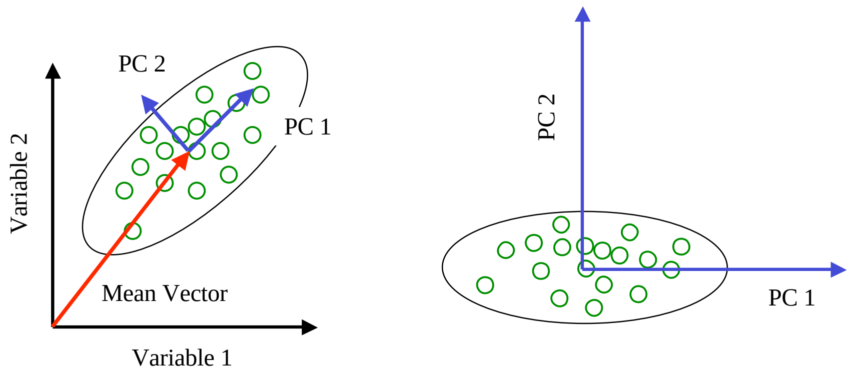
A.M. Belu et. al., "TOF-SIMS Characterization and Imaging of Controlled-Release Drug Delivery Systems, *Anal. Chem.*, **72**(22), pps 5625-5638, 2000.

Principal Components Analysis



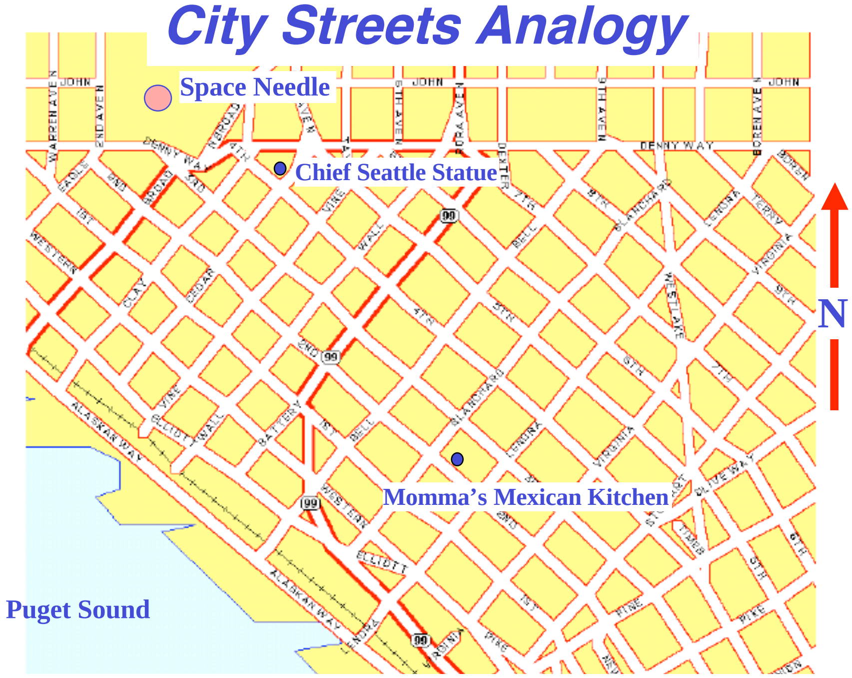
PCA

- Geometry for 2 variables



EIGENVECTOR RESEARCH INCORPORATED 17

City Streets Analogy



PCA Math 1 of 2

For a data matrix **X** with *m* samples and *n* variables (generally assumed to be mean centered and properly scaled), the PCA decomposition is:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + ... + \mathbf{t}_k\mathbf{p}_k^T + ... + \mathbf{t}_q\mathbf{p}_q^T$$

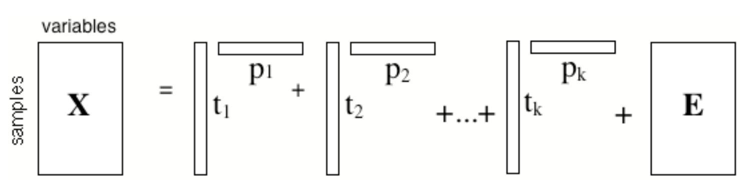
Where $q \leq \min\{m,n\}$, and the $\mathbf{t}_i\mathbf{p}_i^T$ pairs are ordered by the amount of variance captured.

Generally, the model is truncated, leaving some small amount of variance in a residual matrix:

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + ... + \mathbf{t}_k\mathbf{p}_k^T + \mathbf{E} = \mathbf{T}_k\mathbf{P}_k^T + \mathbf{E}$$

EIGENVECTOR RESEARCH INCORPORATED 19

PCA Math 2 of 2



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

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

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

and λ_i are eigenvalues.

Amount of variance captured by $\mathbf{t}_i\mathbf{p}_i^T$ proportional to λ_i.

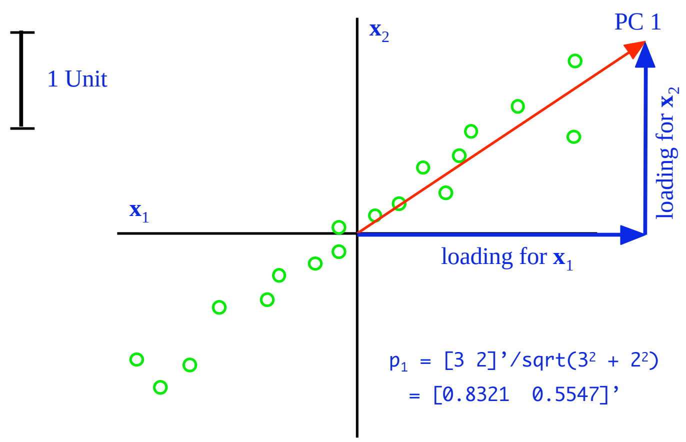
EIGENVECTOR RESEARCH INCORPORATED 20

Properties of PCA

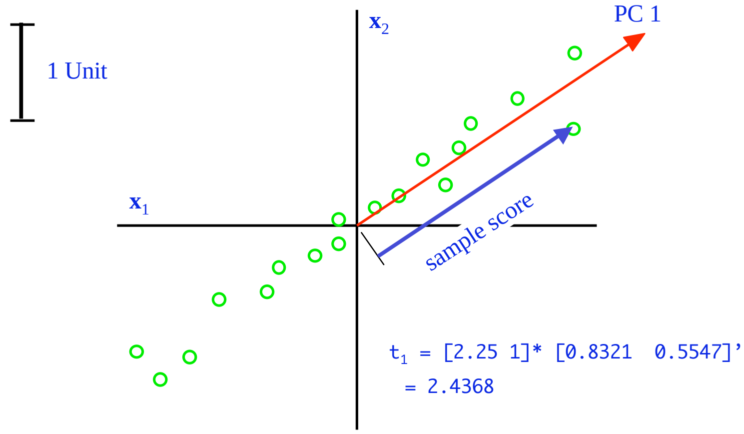
- $\mathbf{t}_i, \mathbf{p}_i$ ordered by amount of *variance captured*
- $\mathbf{t}_i$ or *scores* form an orthogonal set $\mathbf{T}_k$ which describe relationship between *samples*
- $\mathbf{p}_i$ or *loadings* form an orthonormal set $\mathbf{P}_k$ which describe relationship between *variables*

- scores and loadings plots are interpreted in pairs
  - *e.g.* plot $\mathbf{t}_i$ vs sample number and $\mathbf{p}_i$ vs variable number
- it is useful to plot $\mathbf{t}_{i+1}$ vs. $\mathbf{t}_i$ and $\mathbf{p}_{i+1}$ vs. $\mathbf{p}_i$

Variable Loadings, p_i



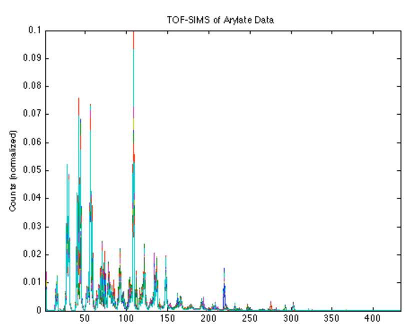
Sample Scores, t_i



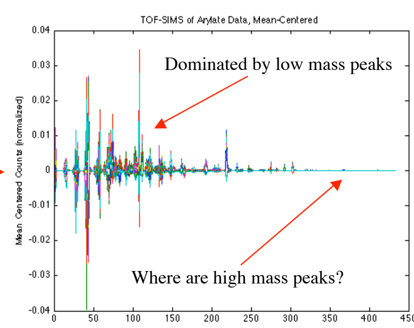
EIGENVECTOR RESEARCH INCORPORATED 23

Arylate Data

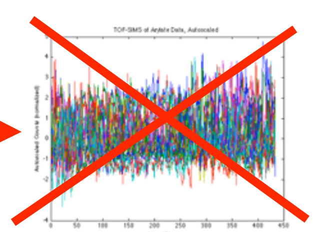
Raw Data



Mean-centered Data



Autoscale?

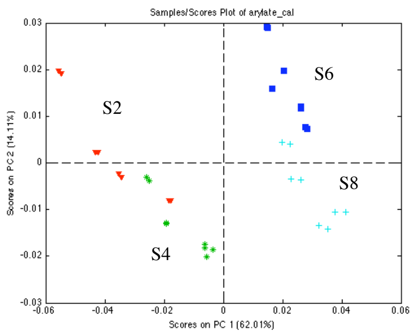


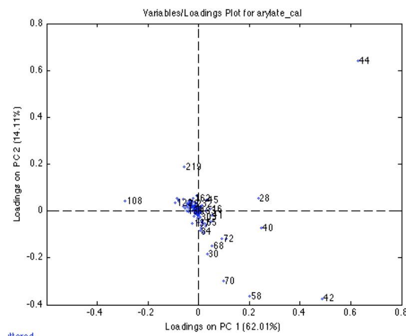
EIGENVECTOR RESEARCH INCORPORATED 24

PCA of Mean-centered Arylate

Percent Variance Captured by PCA Model

Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	8.58e-04	62.01	62.01
2	1.95e-04	14.11	76.13
3	1.65e-04	11.90	88.03
4	6.87e-05	4.97	92.99

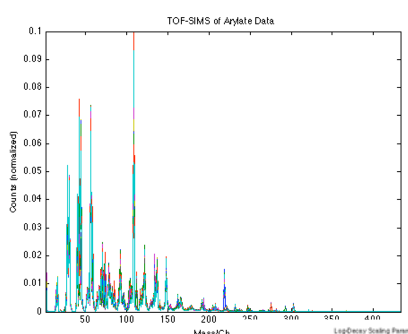


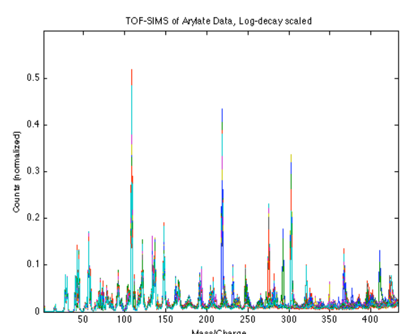


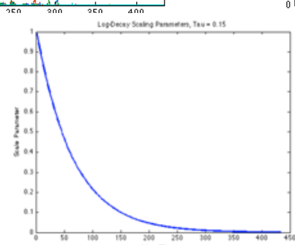
EIGENVECTOR RESEARCH INCORPORATED 25

Log-decay Scaling

Raw Data





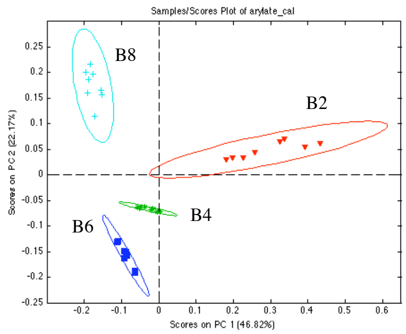


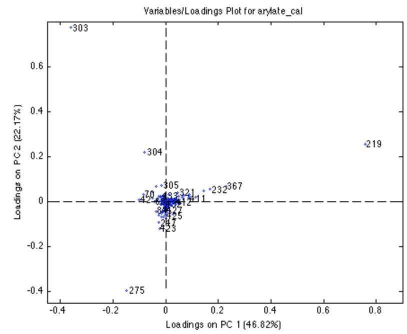
EIGENVECTOR RESEARCH INCORPORATED 26

PCA with Log-decay, MC

Percent Variance Captured by PCA Model

Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	3.47e-02	46.82	46.82
2	1.65e-02	22.17	68.99
3	9.71e-03	13.09	82.08
4	6.50e-03	8.75	90.83





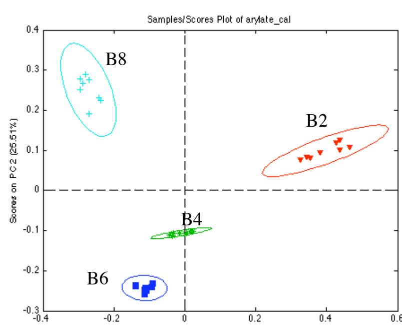
Can we do better? Normalize?

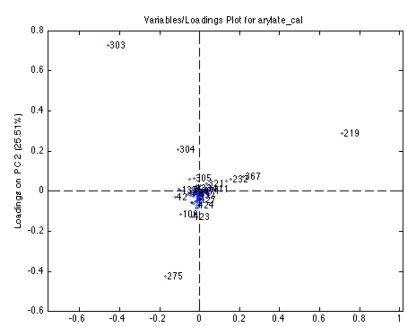
EIGENVECTOR RESEARCH INCORPORATED 27

Log-decay, Normalize, Mean-Center

Percent Variance Captured by PCA Model

Principal Component Number	Eigenvalue of Cov(X)	% Variance Captured This PC	% Variance Captured Total
1	6.39e-02	43.77	43.77
2	3.72e-02	25.51	69.29
3	1.69e-02	11.59	80.88
4	1.40e-02	9.58	90.46

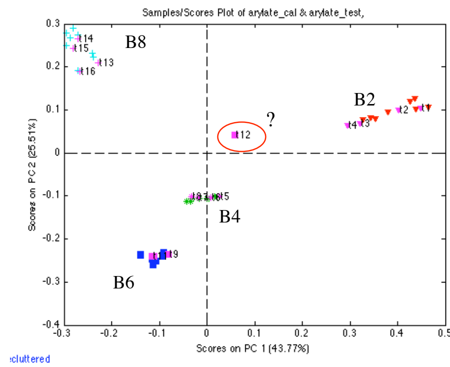


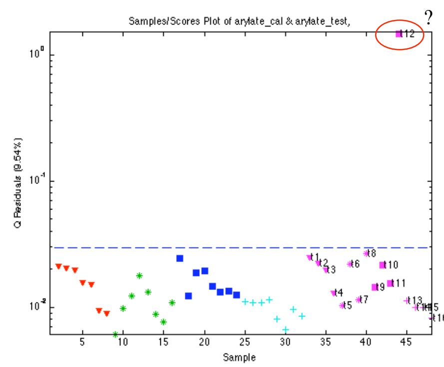


Decluttered

EIGENVECTOR RESEARCH INCORPORATED 28

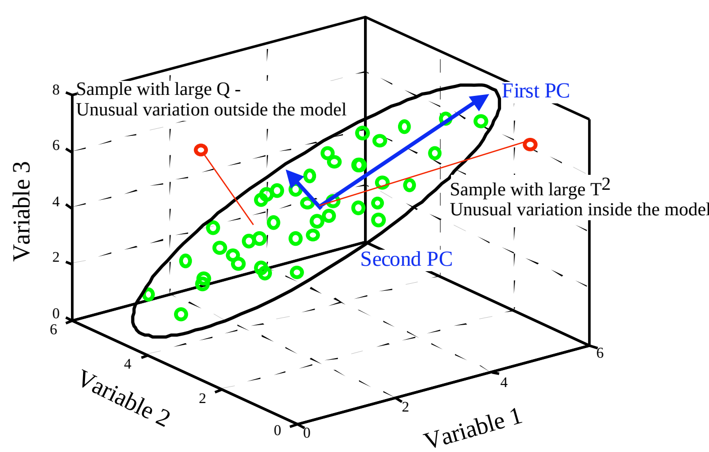
How Does it Work on the Test Set?





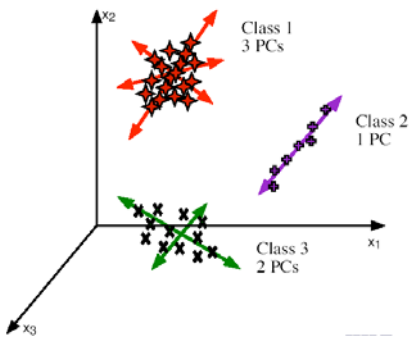
Check residuals!

Geometry of Q and T^2



Supervised Pattern Recognition

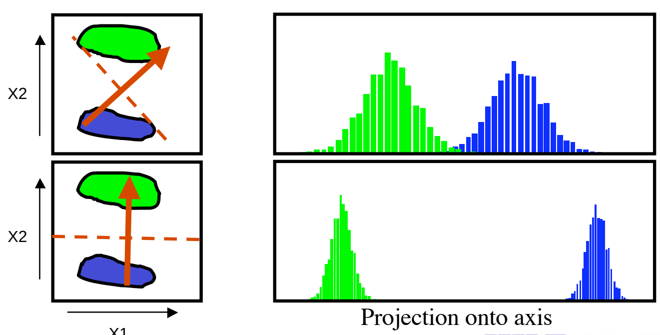
- A single PCA model worked fine to visually classify arylates for backbone length
- PCA models could be built of each class (SIMCA)
- Fairly obvious this would work well



31

Apply SIMCA to Arylate for Sidechain?

- Doesn't work because major variation in spectra (with this scaling) due to backbone, not side chain
- Try discriminant analysis instead



32

Partial Least Squares Discriminant Analysis (PLS-DA)

- Use PLS regression to determine axis to project data on that discriminates between classes
  - choose axis so individual distributions are narrow
  - choose axis so centers of distributions are far apart
- PLS is factor-based model of data therefore more stable with high collinearity.
- Will automatically attempt to identify directions of interest!

PLS-DA for Sidechain Length

Calibration and test samples shown

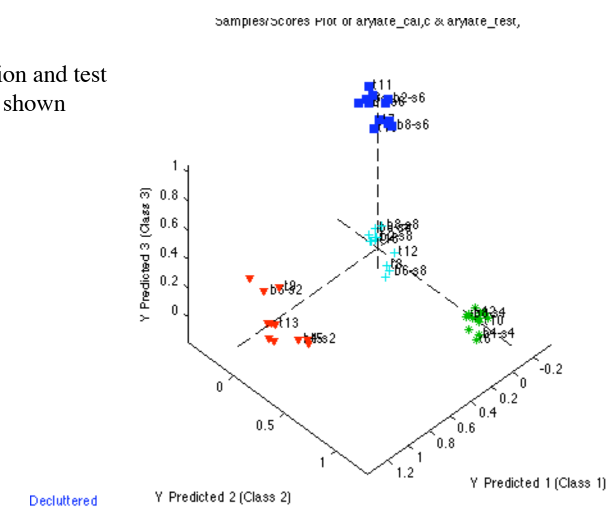
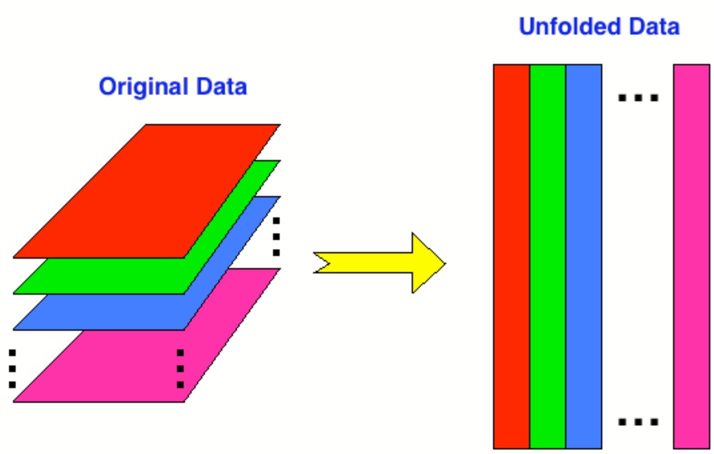


Image PCA

- SIMS images contain complete spectra for each pixel
- Use PCA to condense information from all channels down
- Use “scores” instead of single channels

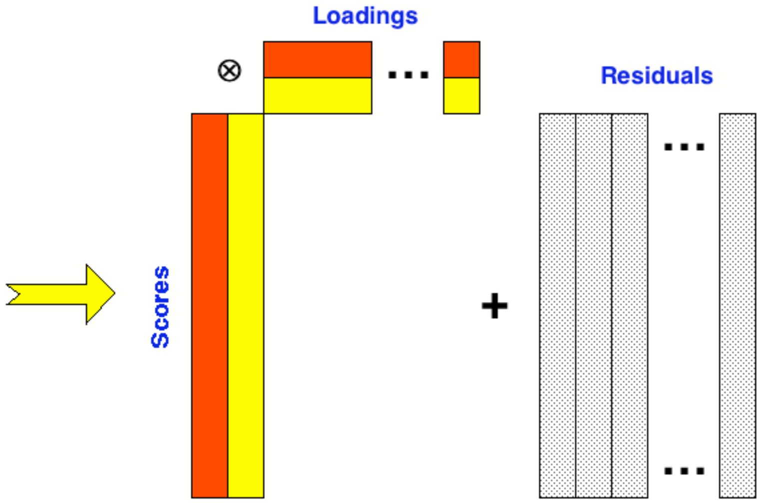
EIGENVECTOR RESEARCH INCORPORATED 35

Matricizing or Unfolding



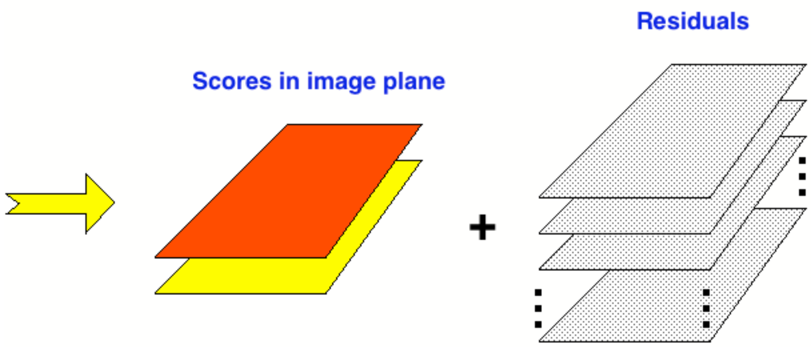
EIGENVECTOR RESEARCH INCORPORATED 36

Perform PCA on Unfolded Data



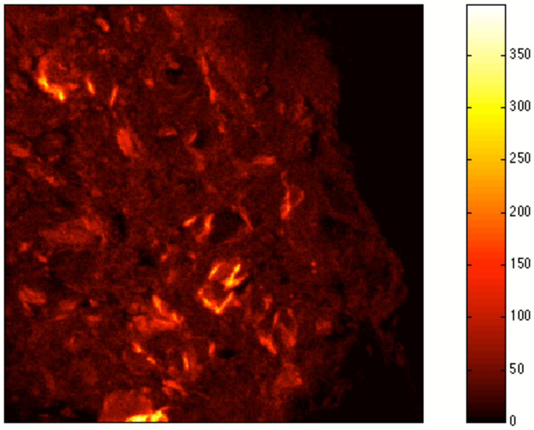
EIGENVECTOR RESEARCH INCORPORATED 37

Refold Results from PCA



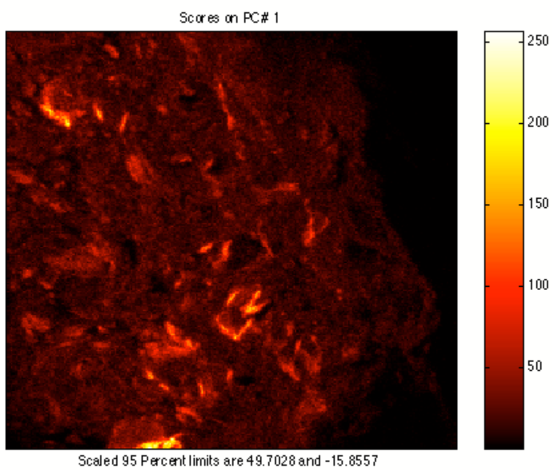
EIGENVECTOR RESEARCH INCORPORATED 38

Total Ion Image of Bead



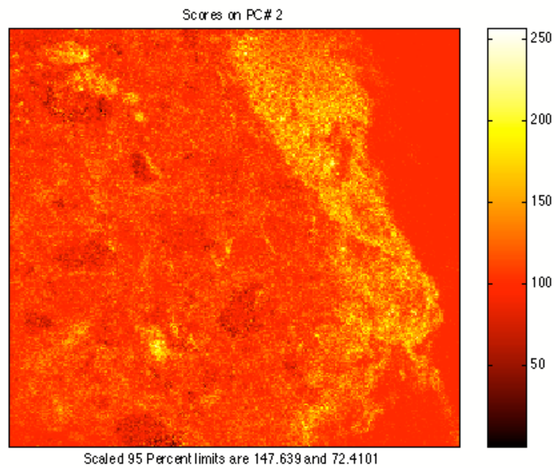
39

Scores on First PC



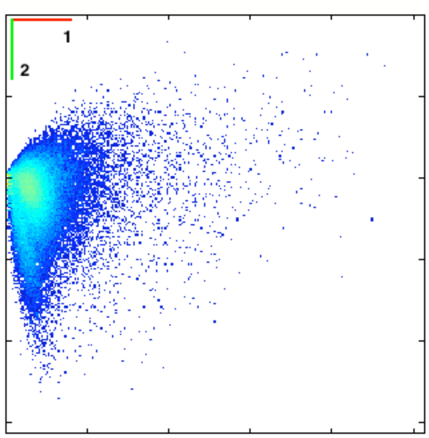
40

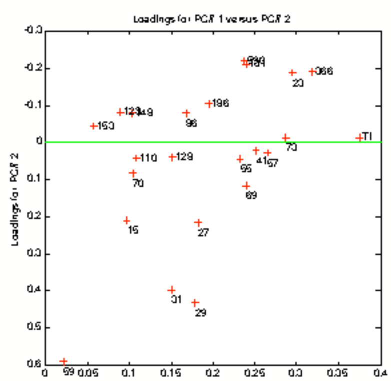
Scores on Second PC



EIGENVECTOR RESEARCH INCORPORATED 41

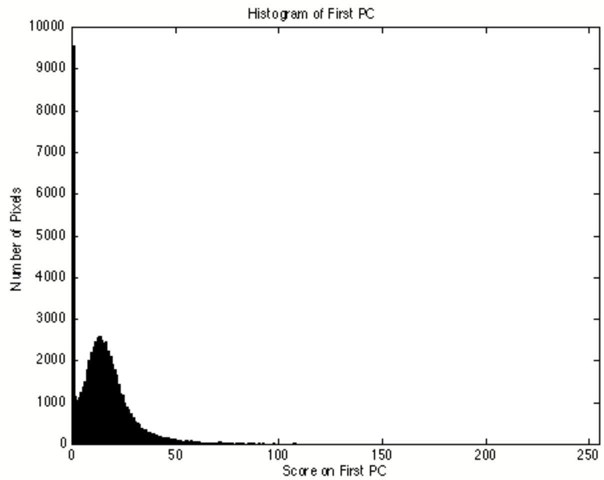
Scores and Loads on Second vs. First PC





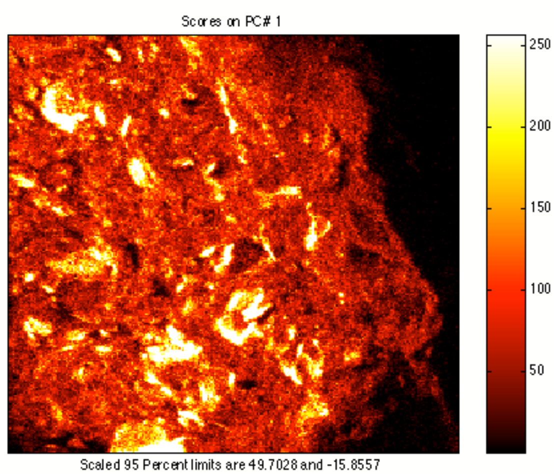
EIGENVECTOR RESEARCH INCORPORATED 42

Problem: Not Much Contrast!



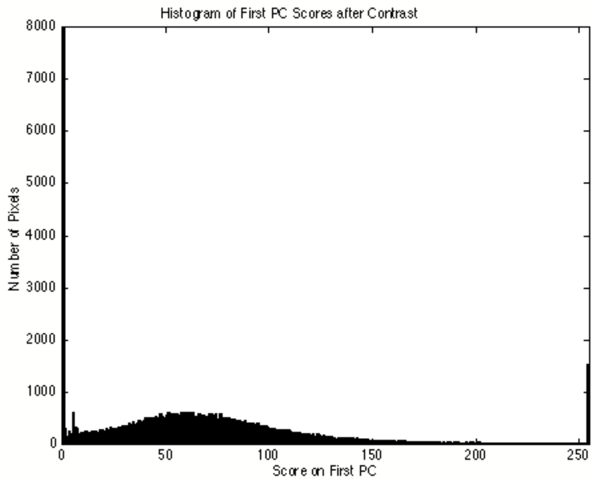
EIGENVECTOR RESEARCH INCORPORATED 43

Contrast Enhanced Scores on PC 1



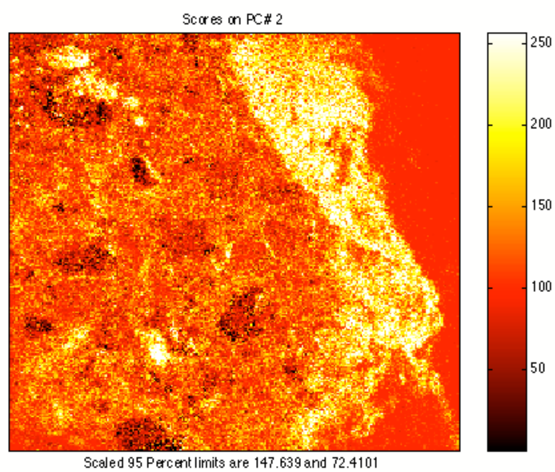
EIGENVECTOR RESEARCH INCORPORATED 44

Histogram of PC1 Scores Afer Contrast Enhancement



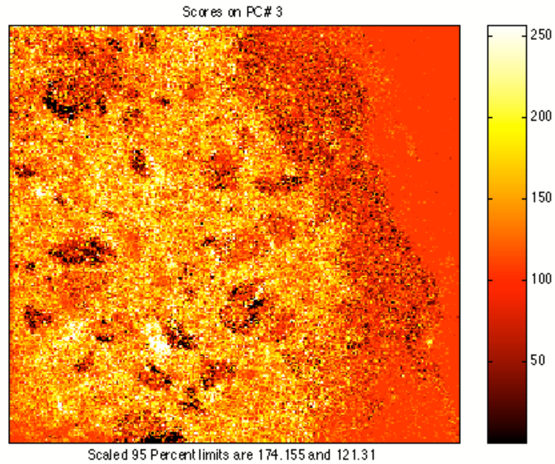
EIGENVECTOR RESEARCH INCORPORATED 45

Contrast Enhanced Scores on PC 2



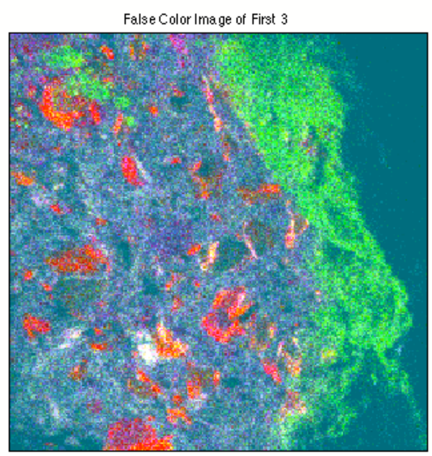
EIGENVECTOR RESEARCH INCORPORATED 46

Contrast Enhanced Scores on PC 3



EIGENVECTOR RESEARCH INCORPORATED 47

Contrast Enhanced False Color Image



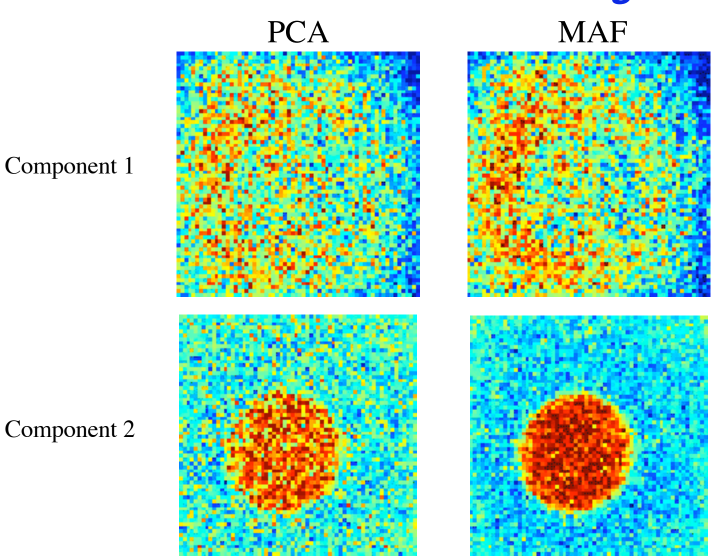
EIGENVECTOR RESEARCH INCORPORATED 48

Maximal Autocorrelation Factors (MAF)

- Regular image PCA does not take any spatial correlations into account, just captures variance
- MAF finds factors which capture large amounts of variance and produce correlated scores in the image plane
- Result is that features with large spatial correlations move up in model

EIGENVECTOR RESEARCH INCORPORATED 49

PVA Image Data
PCA vs. MAF Score Images



MCR Objective

- Decompose a data matrix into chemically meaningful factors
  - pure analyte spectra
  - pure analyte concentrations
- Easy to interpret
  - provides chemically / physically meaningful information
  - caveats:
    - rotational and multiplicative ambiguity
    - use of constraints

EIGENVECTOR RESEARCH INCORPORATED 51

MCR

- Based on the classical least squares (CLS) model, attempt to estimate **C** and **S** given **X**:

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

where

X is a MxN matrix of measured responses,

C is a MxK matrix of pure analyte contributions,

S is a NxK matrix of pure analyte spectra, and

E is a MxN matrix of residuals.

Also called Self-modeling Mixture Analysis

EIGENVECTOR RESEARCH INCORPORATED 52

Alternating Least Squares

- How can we improve estimates of **S** and **C**?
- Given initial guess $\mathbf{S}_0$ (or $\mathbf{C}_0$)...

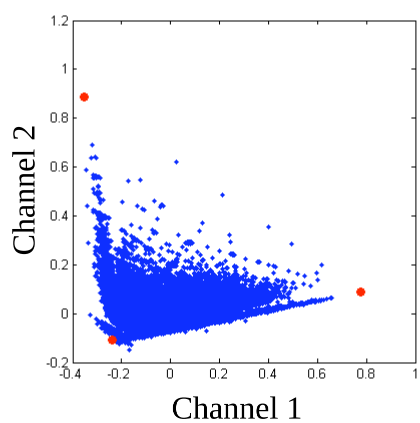
  $\mathbf{C}_i = \mathbf{X}\mathbf{S}_{i-1}(\mathbf{S}_{i-1}^{\mathrm{T}}\mathbf{S}_{i-1})^{-1}$

  $\mathbf{S}_i = (\mathbf{C}_i^{\mathrm{T}}\mathbf{C}_i)^{-1}\mathbf{C}_i^{\mathrm{T}}\,\mathbf{X}$
- Iterate until convergence (ALS)
  - Usually constrained such that **C**>0 and **S**>0
  - and each $\mathbf{s}_k^T\mathbf{s}_k=1$

EIGENVECTOR RESEARCH INCORPORATED 53

Initial Estimate

- Try to find "extreme" samples/pixels
- Or look for "extreme" variables



MCR (ALS) on TOF-SIMS Image

- Non-negative constraints on both C and S
- Initialize with pure samples (i.e. pixels)
- Recover 6 interpretable spectra and concentration profiles
- Showing Score Images – image was unfolded with each pixel as a separate sample then the scores are re-folded to form images

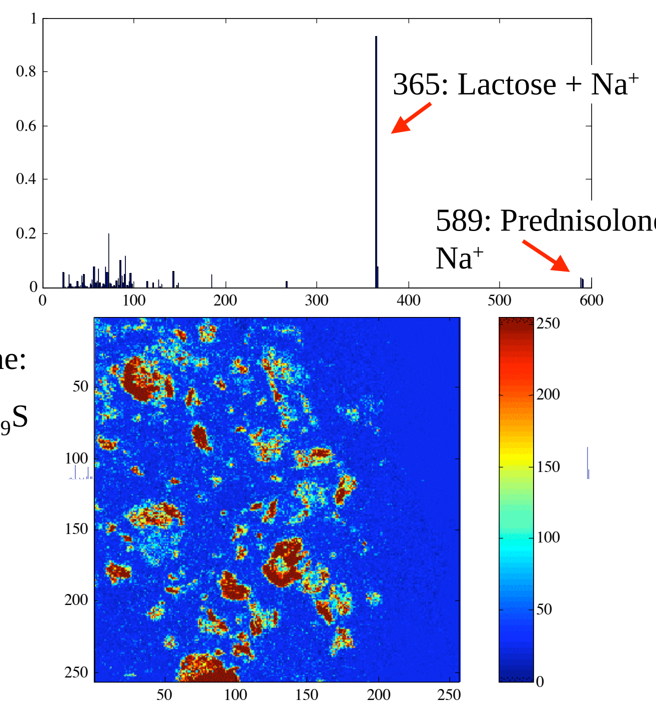
EIGENVECTOR RESEARCH INCORPORATED 55

Prednisolone:

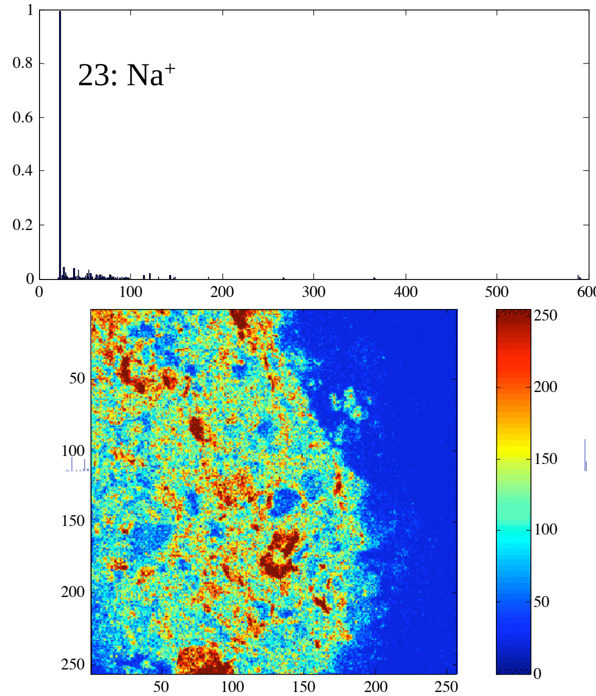
$C_{21}H_{31}NaO_9S$

Lactose:

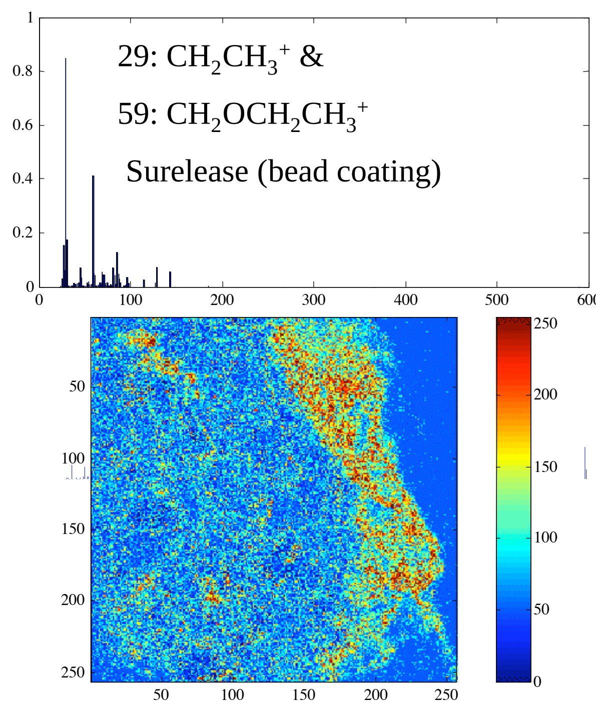
$C_{12}H_{22}O_{11}$



56



57

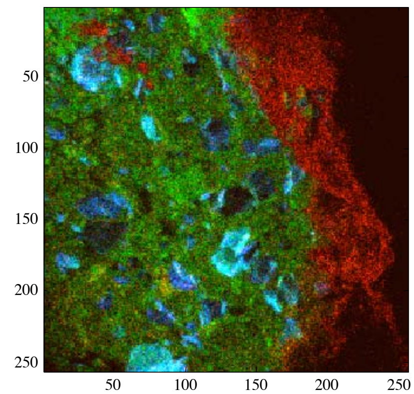


58

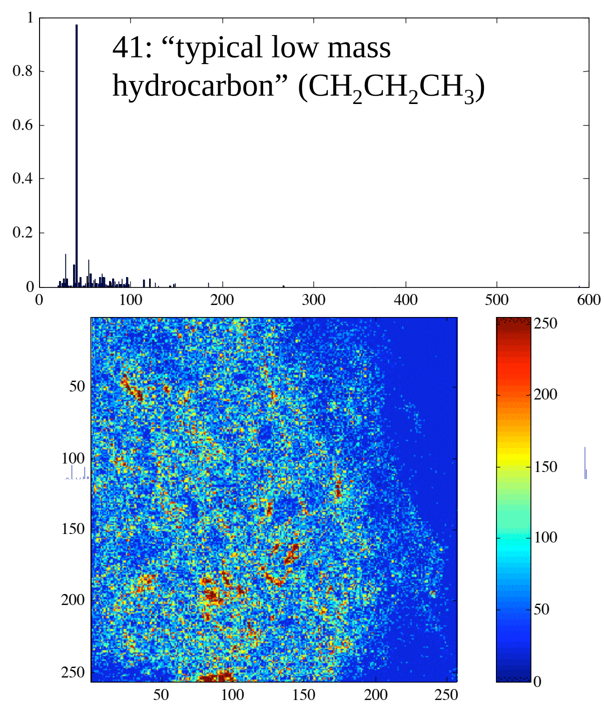
RGB “Chemical” Image

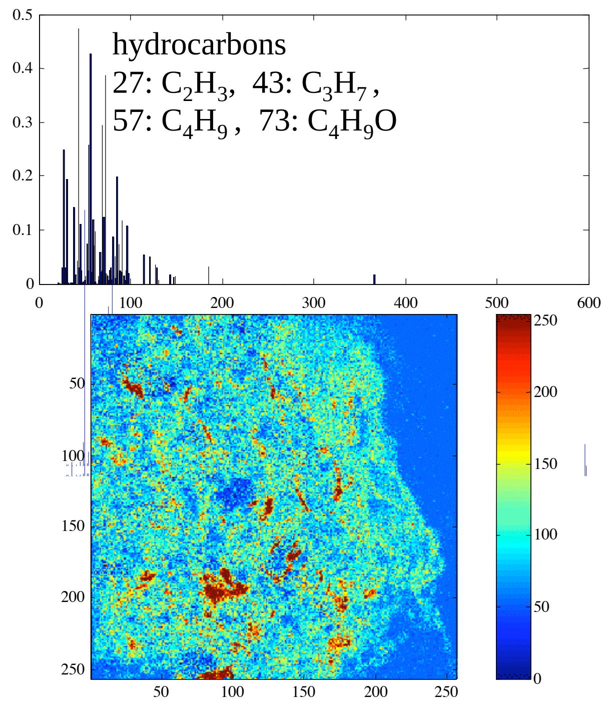
Red: Surelease (bead coating)
Green: Na
Blue: Prednisolone (drug)

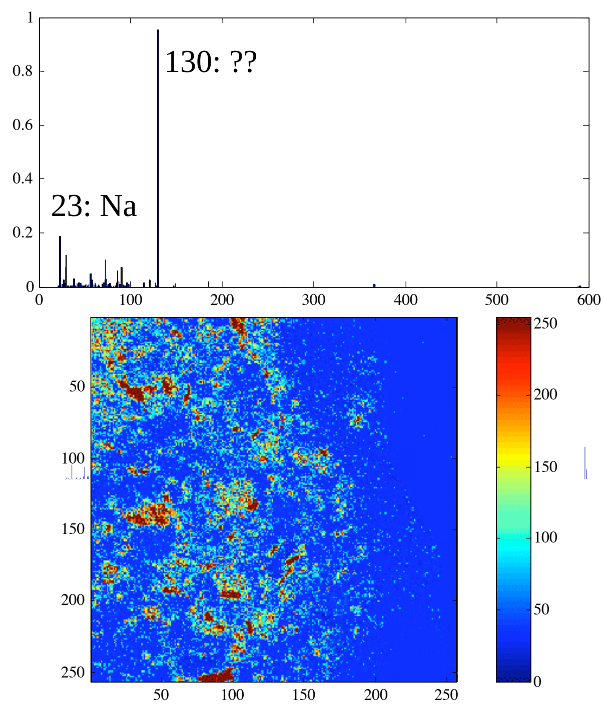
only 3 of 6 factors extracted are shown



EIGENVECTOR RESEARCH INCORPORATED 59

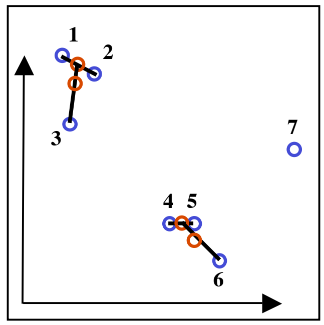



hydrocarbons
27: C_2H_3, 43: C_3H_7,
57: C_4H_9, 73: C_4H_9O


130: ??
23: Na

k-Means Agglomerative Clustering

- Samples are paired with another sample or a cluster one-at-a-time
- Position of each cluster is mean of all samples in cluster.
- Recalculation of distance can take a long time with lots of samples

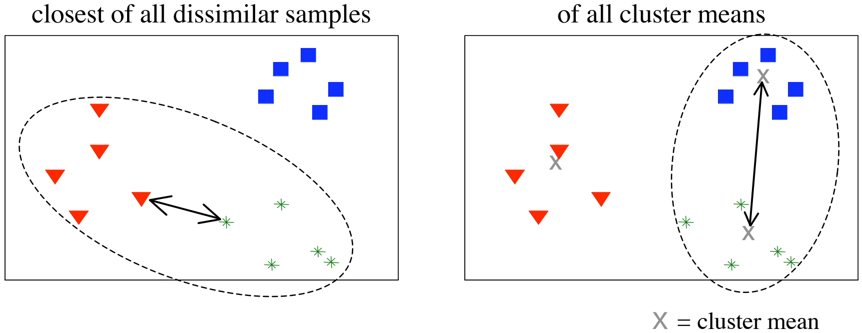


KNN vs. K-Means

Two clusters are grouped together when…

KNN
…two of their members are the closest of all dissimilar samples

K-Means
…the cluster means are the closest of all cluster means



Note: these rules apply even when one of the "groups" is a single sample in a group of its own.

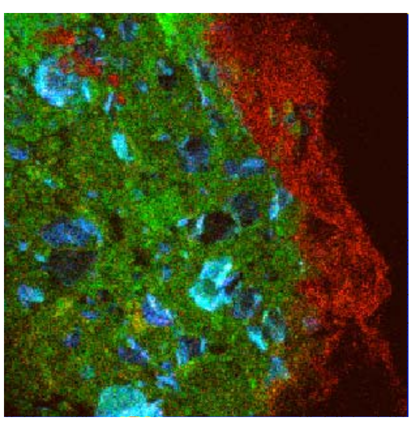
k-Means Partitional Clustering

- Choose k samples as cluster "targets"
  - random selection of samples
  - "pure samples": choose samples on outside of data (furthest from all other samples)
- Classify all samples into one of those k clusters.
- Calculate mean of each cluster's samples
- Repeat classification and cluster means until no samples are re-classed after mean recalculation.
- Much faster, but dependent on initial guess of samples

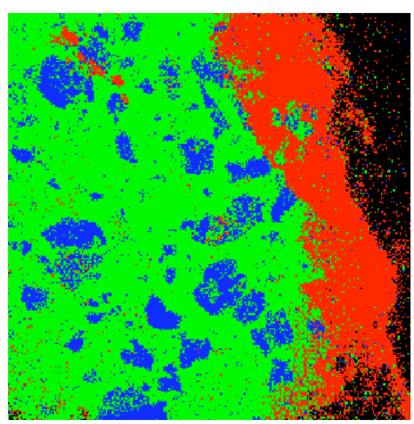
65

Avicel by k-means Clustering

False-color MCR Results



Pure Pixel Clusters



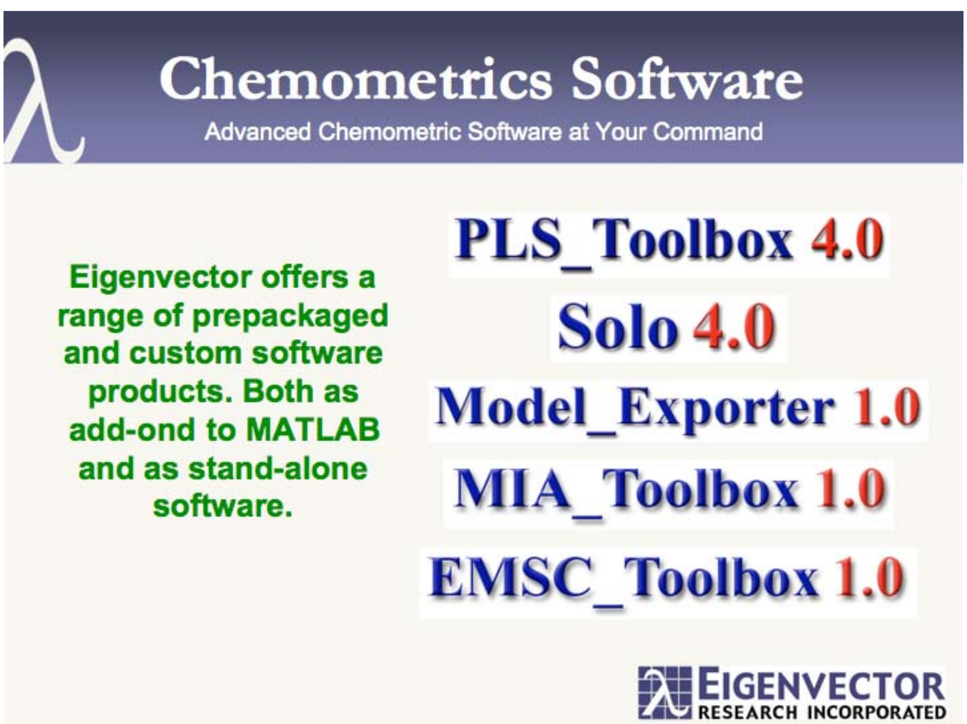
(3 clusters)

66

Why Multivariate and Factor Based Methods?

- Noise filtering
- Selectivity enhancement
- Interpretation
- It's a multivariate world!

EIGENVECTOR RESEARCH INCORPORATED 67



Resources

- Books
- *Chemometrics*, M.A. Sharaf, D.L. Illman and B.R. Kowalski, Wiley-Interscience (1986) ISBN 0-471-83106-9
- *Multivariate Analysis*, K.V. Mardia, J.I. Kent and J.M. Bibby, Academic Press, (1979) ISBN 0-12-471252-2
- *Multivariate Calibration*, H. Martens and T. Næs, John Wiley & Sons Ltd. (1989) ISBN 0-471-90979-3
- *Chemometrics: a textbook*, D.L. Massart et al., Elsevier (1988) ISBN 0-444-42660-4
- *Chemometrics: A Practical Guide*, K.R. Beebe, R.J. Pell, M.B. Seasholtz, Wiley (1998) ISBN 0-471-12451-6
- *Multivariate Data Analysis In Practice*, Kim H. Esbensen, CAMO ASA (2000), ISBN 82-993330-2-4
- *A user-friendly guide to Multivariate Calibration and Classification*, T. Næs, T. Isaksson, T. Fearn, T. Davies, NIR Publications(2002), ISBN 0-9528666-2-5
- *Multivariate Image Analysis*, Paul Geladi and Hans Grahn, Wiley (1996), ISBN 0-471-93001-6
- *Multivariate Analysis of Quality: An Introduction*, H. Martens and M. Martens, Wiley (2001), ISBN 0-471-97428-5
- Journals
- Journal of Chemometrics
- Chemometrics and Intelligent Laboratory Systems
- Analytical Chemistry
- Analytica Chemica Acta
- Applied Spectroscopy
- Critical Reviews in Analytical Chemistry
- Journal of Process Control
- Computers in Chemical Engineering
- Technometrics
- ....

EIGENVECTOR RESEARCH INCORPORATED 69