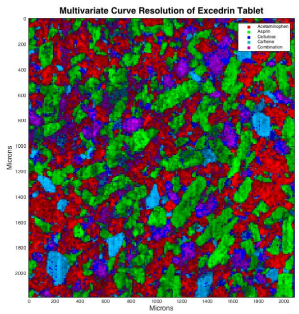


Introduction to Hyperspectral Image Analysis (aka MIA)

EIGENVECTOR RESEARCH INCORPORATED

Table of Contents

EIGENVECTOR RESEARCH INCORPORATED

Table of Contents (cont.)

EIGENVECTOR RESEARCH INCORPORATED

Resources

- *Hyperspectral Image Analysis*, eds. P. Geladi and H. Grahn, Wiley (2007), ISBN 978-0-470-01086-0
- *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

- Journal of Chemometrics
- IEEE Trans. on Geosci. and Remote Sensing
- 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
- ....

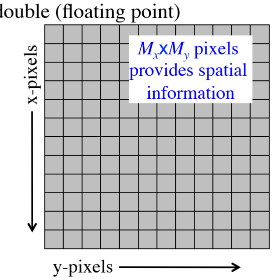
EIGENVECTOR RESEARCH INCORPORATED

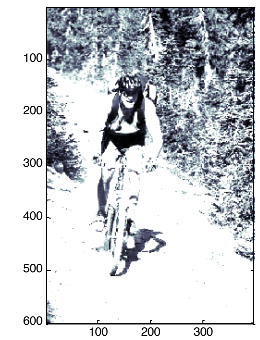
Course Materials

- These slides
- PLS_Toolbox and MIA_Toolbox or Solo+MIA
- Data sets
  - From DEMS folder (distributed with software)
    - EDS Wire Alloy,
  - From EVRIHW folder (additional data sets)
    - Nuts3.jpg, Mississippi, Excedrin_small, PVA, bananas

Univariate Image

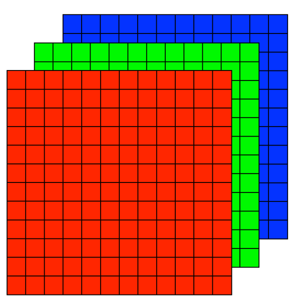
- Grey scale
  - each pixel is an number defining an intensity level e.g.,
    - integer (0 to 255) unsigned 8-bit
    - integer (0 to 4095)
    - double (floating point)





Multivariate Image
(3 Variables)

- Red/Green/Blue (RGB) (e.g. JPEG)
  - each layer defines color intensity level
  - much more information-rich



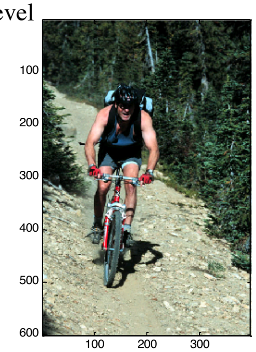
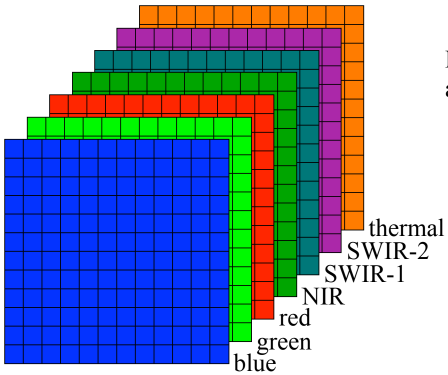


Image Analysis

- Many methods have been developed to examine the spatial structure w/in an image
  - the methods recognize spatial patterns within an image
    - based on the light / dark contrast and continuity of regions
  - edge detection, image sharpening, wavelets
  - particle size distributions, machine vision, medical applications, security, …
- MIA has been traditionally applied to the spectral dimension first followed by spatial analysis
  - some methods that examine both are appearing

Multivariate Image
(4-10 Variables)

- Measure at several wavelengths (e.g., Landsat)



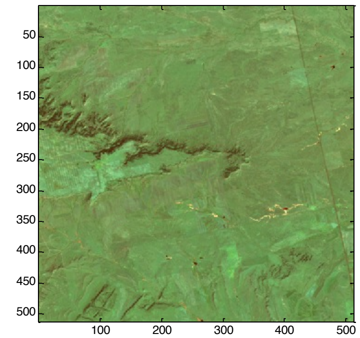
How should we display a seven variable image?

EIGENVECTOR RESEARCH INCORPORATED

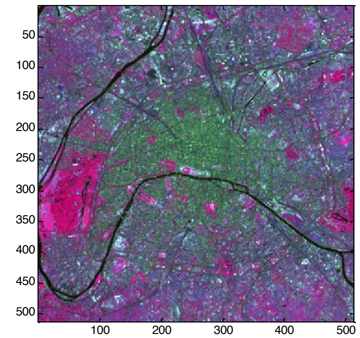
Multivariate Image
(4-10 Variables)

- Choose 3 of 7 (Landstat)

Montana (blue/SWIR-1/thermal)



Paris (NIR/blue/SWIR-1)*

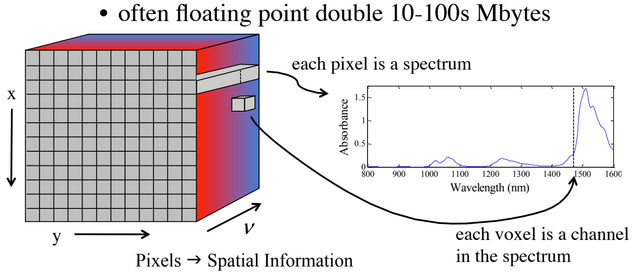


*contrast enhanced

EIGENVECTOR RESEARCH INCORPORATED

Hyperspectral Image
(>10 Variables)

- Spectrum at each pixel
  - could be 100-1000s of variables
  - often floating point double 10-100s Mbytes



Pixels → Spatial Information
Spectral → Chemical Information

EIGENVECTOR RESEARCH INCORPORATED

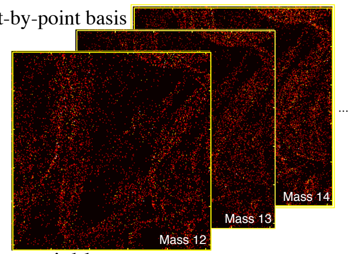
Multivariate Images

- 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 (e.g, spectroscopy).
- Chemical system(s) of interest include
  - microscopic, medical, machine vision, process monitoring crystallization, stand-off and remote sensing, …
  - vapors, liquids, solids (or combination)
  - visible, infra-red, Raman, mass spectroscopy, …

EIGENVECTOR RESEARCH INCORPORATED

Physics of Measurement

- Point scanning
  - spectra measured on a point-by-point basis
  - secondary-ion mass spec
  - atomic force microscopy
  - surface Raman
- Line scanning
  - push broom
- Focal plane array
  - images can be acquired very quickly

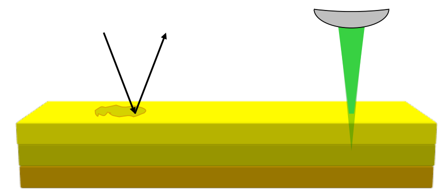


EIGENVECTOR RESEARCH INCORPORATED

Volumetric Analysis Techniques

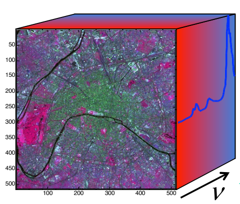
- Confocal Wavelength Resolved Imaging
- Surface Ablation Techniques

- Produces multivariate data in *3-dimensional space*



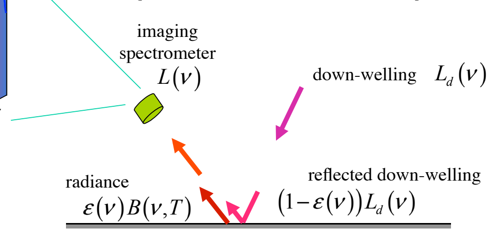
EIGENVECTOR RESEARCH INCORPORATED

Standoff and Remote Sensing



- Detection of residues on, and under, surfaces at standoff distances using hyperspectral imaging

$$L(\nu)=\left\{\varepsilon(\nu)B(\nu,T)+\left(1-\varepsilon(\nu)\right)L_d(\nu)\right\}\tau(\nu)$$



EIGENVECTOR RESEARCH INCORPORATED

Simple Image Analysis Tools

- TrendTool – Univariate Data Investigation
  - Analyze multivariate data using simple univariate measurements
- Image Manager – Data Manipulation and Analysis
  - Concatenating / Manipulating (e.g. rotation) Images
  - Particle Analysis
- Image Exploration Tools
  - Cross-section, Drill, and Magnification
- Preprocessing

EIGENVECTOR RESEARCH INCORPORATED

TrendTool

- Display results of univariate calculations on multivariate data
  - Signal at given variable
  - Integrated signal across range of variables
  - Peak position
  - Peak width
- With or without baselines
- Ratio of measurements

Opening TrendTool

Workspace Browser

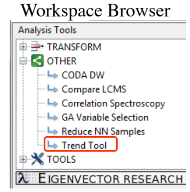
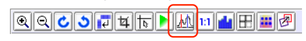
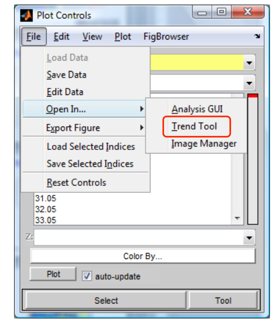


Image Manager Toolbar



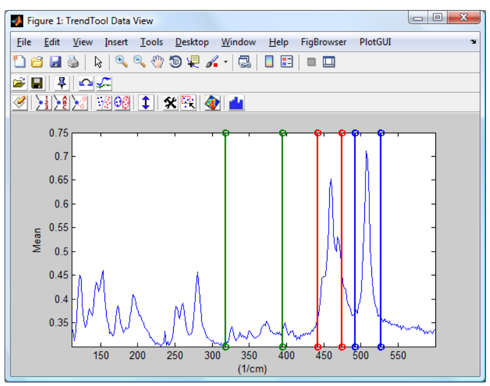
Plot Controls Window



TrendTool Windows: Data View

Use Data View to:
- Set analysis markers
- Choose analysis mode
- Select references and baseline points

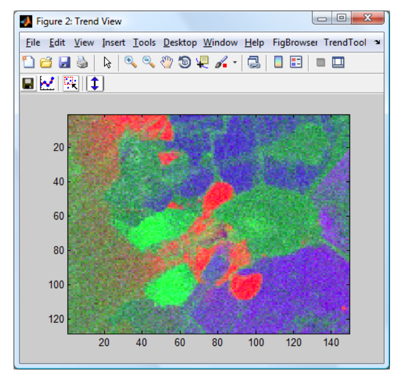
Hints:
- Right-click white space to set marker or use toolbar button
- Drag markers to move
- Right-click markers to change types
- Use toolbar to save or load marker sets



TrendTool Windows: Trend View

Results displayed in Trend View
- Single marker displays with false-color
- Multiple markers display in RGB

Toolbar Buttons:
- autoscale image
- select pixels to display in Data View
- save or spawn plot of results (respectively)



TrendTool Analysis Modes

- **Height** – gives response at position (single marker)
- **Area** – gives integrated response between markers
- **Position** – gives position of peak response between markers
- **Width** – gives full width at half height between markers

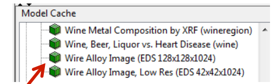
"Add Reference" to subtract a single point baseline. Convert reference to baseline (via right-click) to do two-point linear baseline.

"Normalize to Region" to normalize all regions to the response of the selected region.

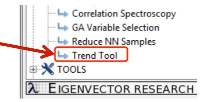
EIGENVECTOR RESEARCH INCORPORATED

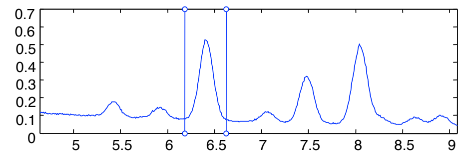
TrendTool Example

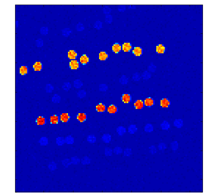
Example: "wires" dataset
Energy Dispersive X-Ray Spectroscopy (EDS)
Image of wires composed of different alloys.



•Workspace Browser: Model Cache > Demo Data
•Drag "Wire Alloy Image" to TrendTool in Other Analysis Tools
•Use TrendTool to look at various peaks (right-click peak to change to peak type)

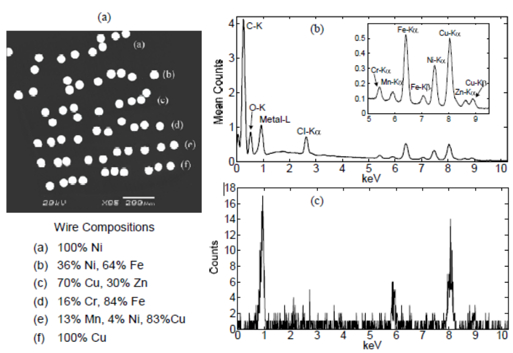






EIGENVECTOR RESEARCH INCORPORATED

Energy dispersive spectrometry (EDS)



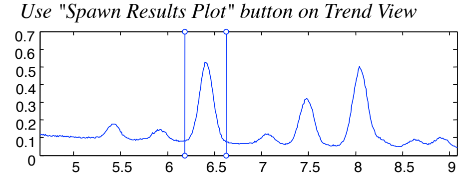
M.R. Keenan, Multivariate Analysis of Spectral Images Composed of Count Data, In: H. F. Grahn, P, Geladi (eds.), Techniques and Applications of Hyperspectral Image Analysis, pp. 89-126, Wiley & Sons, 2007

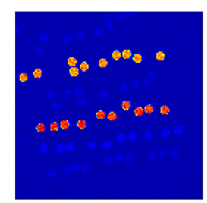
EIGENVECTOR RESEARCH INCORPORATED

Image Exploration

- Cross-section Tool – Transect of spatial dimension
- Drill Tool – Profile through variables of image
- Magnification Tool – Enhance spatial visibility

Example: "wires" dataset using TrendTool to look at one or more peaks…
Use "Spawn Results Plot" button on Trend View

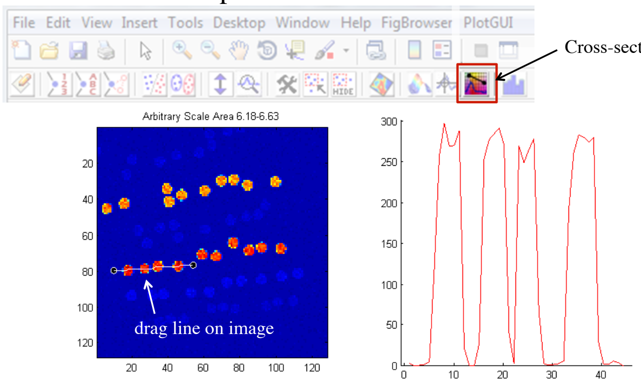




EIGENVECTOR RESEARCH INCORPORATED

Cross-Section Tool

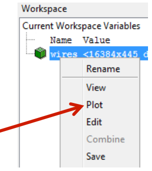
- Transect of spatial dimensions



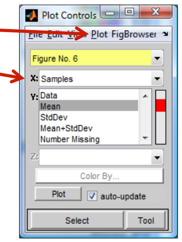
Cross-section Tool

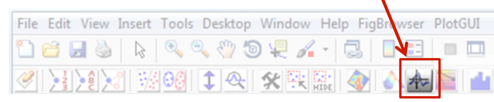
Drill Tool

- Display of data under a given point

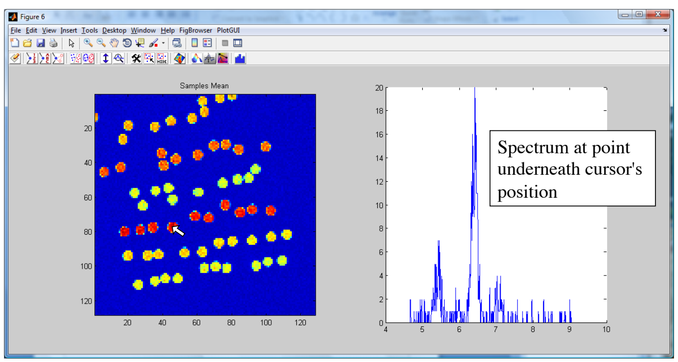


Example: "wires" dataset
Workspace Browser: right-click wires > Plot
Select menu: Plot > Data Summary
Select x-axis menu control: "Samples"
Select toolbar button: "Open/Close Drill Axis"





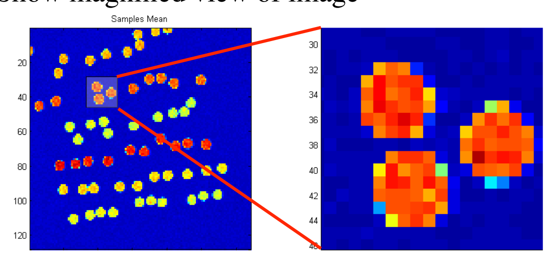
Drill Tool

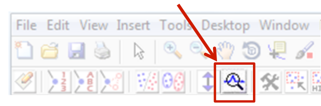


Double-click to view multiple spectra

Magnification Tool

- Show magnified view of image

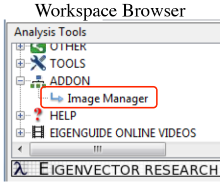




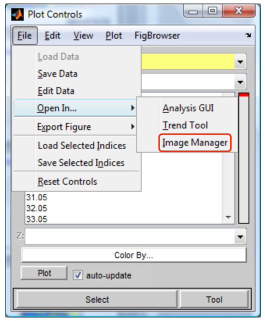
Drag region around to move magnifying glass
Make selections/etc on either image

Opening Image Manager

Workspace Browser



Plot Controls Window



Plot Toolbar

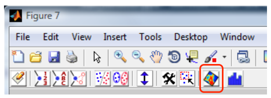


Image Manager Overview

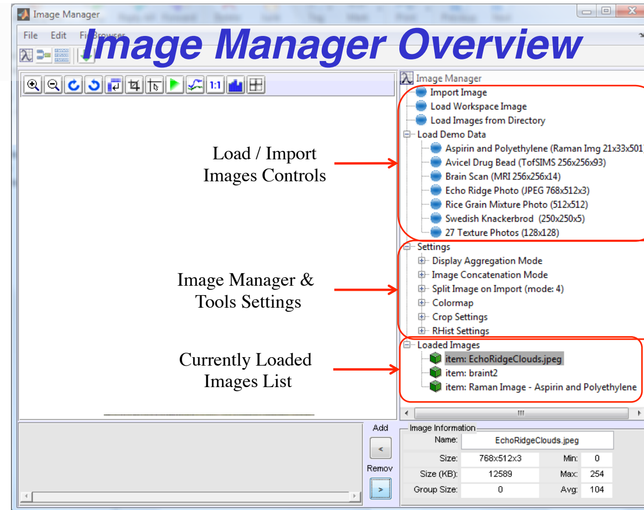


Image Manager Overview

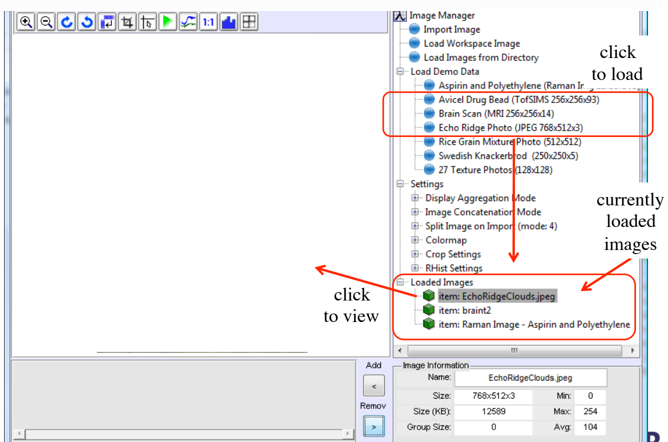
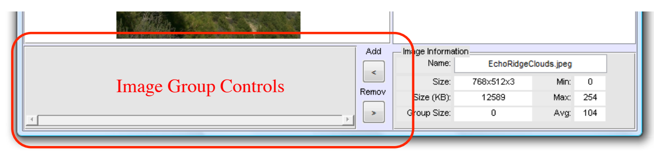


Image Groups

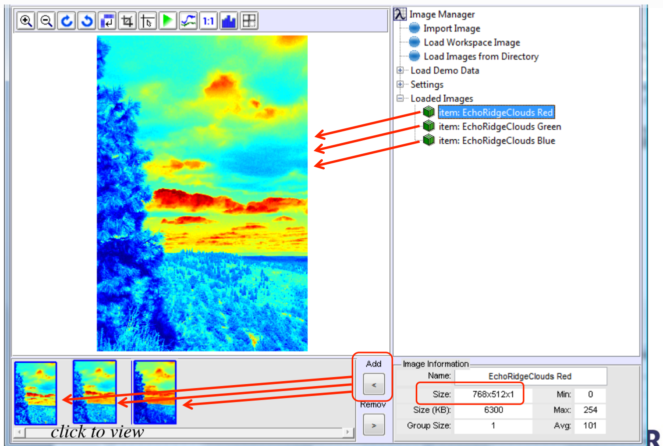
Grouping allows you to:

- Combine images into a single DataSet for analysis
- Apply a univariate operation (rotate, crop, etc) to all images

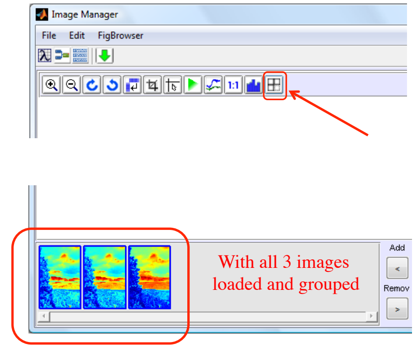


Example: combining three slabs of RGB image

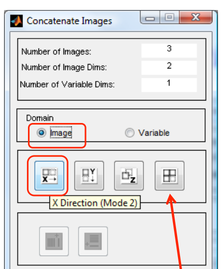
Image Groups

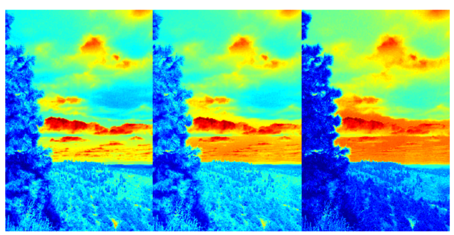


Concatenating Images



Concatenating Images: Spatial Domain

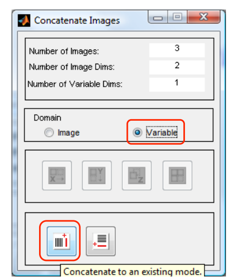




(768 x **1536**) x 1

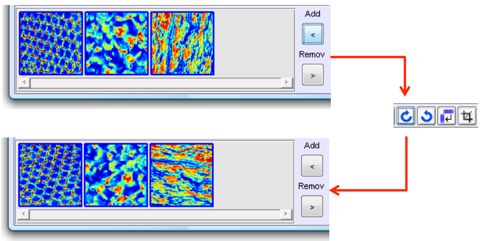
X, Y, Z, or tile…

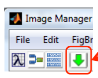
Concatenating Images: Variable Domain



(768 x 512) x **3**

Group Manipulation Example: Rotation





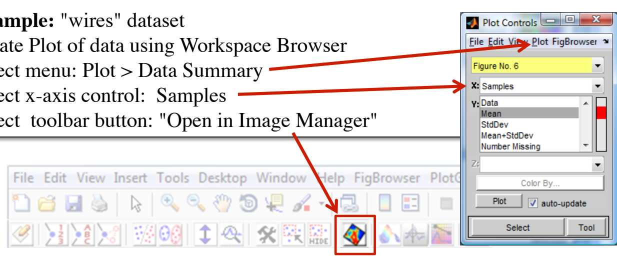
Hint: to apply an action to only ONE image, click the "Apply Changes to Image Group" button until only one thumbnail is outlined in the image group pane.

Particle Analysis

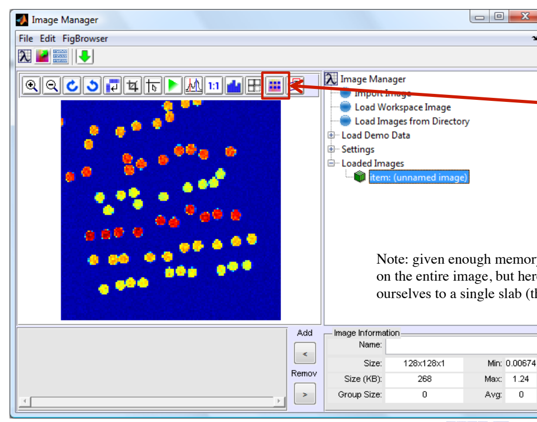
- Identify isolated regions (particles) in an image and give statistics on individual particles.
- Screen out particles and/or background.
- Create models based on particle statistics.
  - Particle outlier models (e.g. identify unusual particles)
  - Inferential models (e.g. drug activity based on particle statistics)
- Based on long-established ImageJ platform.

Particle Analysis Example

Example: "wires" dataset
Create Plot of data using Workspace Browser
Select menu: Plot > Data Summary
Select x-axis control: Samples
Select toolbar button: "Open in Image Manager"

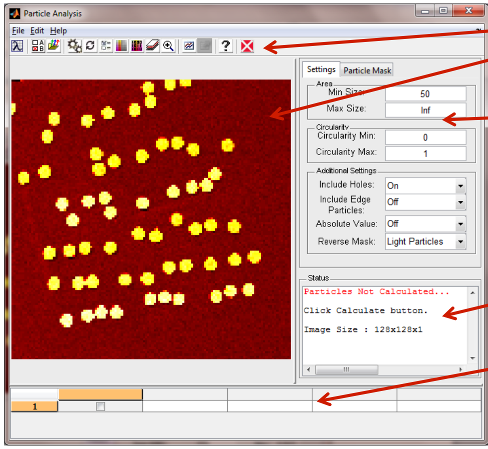


Particle Analysis Example



Select Particle Analysis tool

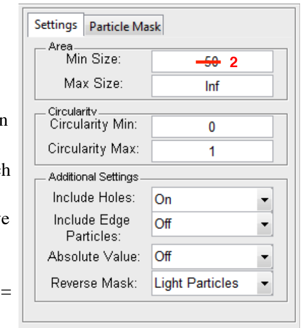
Note: given enough memory, you COULD work on the entire image, but here we're limiting ourselves to a single slab (the mean).



Actions Toolbar
Image with mask
Particle Analysis Settings
Summary of results
Table of located particles

Particle Analysis Settings

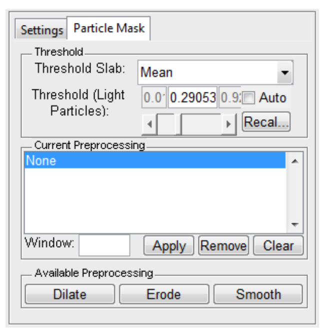
- **Area Min/Max:** Ignore particles with area outside this range.
- **Circularity Min/Max:** Ignore particles outside this range.
- **Include Holes:** On = Include centers of particles even if below threshold.
- **Include Edge Particles:** On = Include particles which touch the edge of the image.
- **Absolute Value:** On = Consider positive and negative deviation from zero as "on" when making mask.
- **Reverse Mask:** Light Particles = Low signal is considered "off" (dark = not particle). Dark Particles = Low signal is considered "on" ("dark image" mode).



Particle Mask Settings

Adjusts which pixels are considered particles

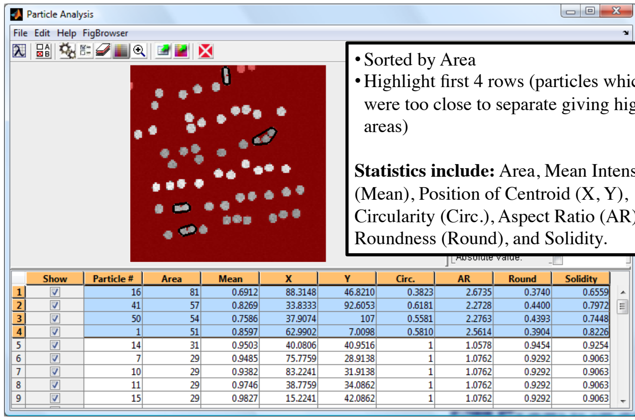
- **Threshold Slab:** For multi-slab images, which image slab is used to mask.
- **Threshold:** Signal level separating particles from background (slider adjusts or "Auto" checkbox does automatic threshold detection.)
- **Preprocessing:** Allows various operations on the binary image mask:
  - **Dilate:** Decrease mask around unmasked regions
  - **Erode:** Increase mask around unmasked regions.
  - **Smooth:** Smooth out noise in mask.



Particle Analysis Example

- On "settings" tab, set Min Size to "2"
- On "Particle Mask" tab, set threshold to "0.4"
- Click "Recalc" button (next to threshold)
- Use Background Color and Grayscale settings to adjust display.
- Select row of table to highlight corresponding particle.
- Select particle in image to highlight corresponding row of table.
- Sort by column using right-click menu.
- Use Export toolbar buttons to send table or image to Analysis.

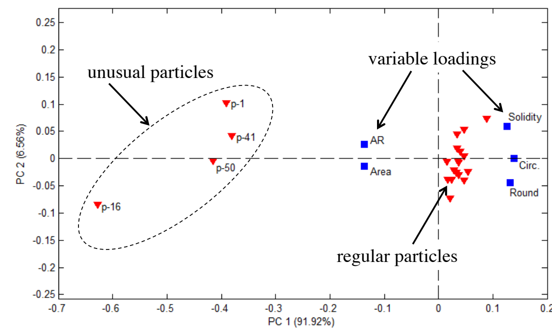
Particle Analysis Example



- Sorted by Area
- Highlight first 4 rows (particles which were too close to separate giving high areas)

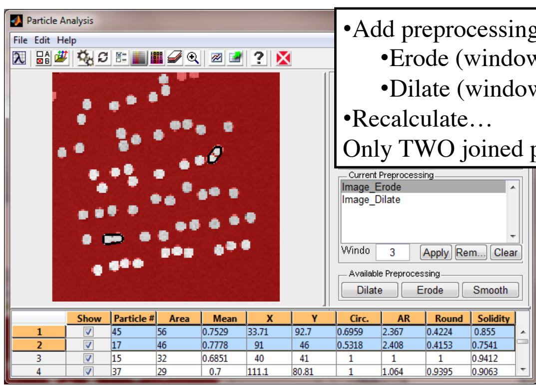
Statistics include: Area, Mean Intensity (Mean), Position of Centroid (X, Y), Circularity (Circ.), Aspect Ratio (AR), Roundness (Round), and Solidity.

PCA of Particle Statistics Biplot of PCs 1 and 2



Autoscaled PCA model with mean intensity (Mean) and centroid (X, Y) variables excluded

Using Preprocessing



- Add preprocessing:
  - Erode (window = 3)
  - Dilate (window = 3)
- Recalculate…

Only TWO joined particles

Image-Oriented Preprocessing

- Image-specific preprocessing operates in pixel-space and are either Intensity or Binary based
- Intensity-Based Image Correction:
  - *Background Subtraction (Flatfield)*: Rolling-ball background subtraction for images.
  - *Min*: Min value over neighboring pixels. (filter out high-value pixels)
  - *Max*: Max value over neighboring pixels. (filter out low-value pixels)
  - *Mean*: Mean value over neighboring pixels. (filter out low/high pixels)
  - *Median*: Median value over neighboring pixels. (robust filter of low/high pixels)
  - *Trimmed Mean*: Trimmed mean value over neighboring pixels.
  - *Trimmed Median*: Trimmed median value over neighboring pixels.
  - *Smooth*: Spatial smoothing for images. (a weighted mean)

Image-Oriented Preprocessing

- Binary-Based Image Correction
  - *Dilate*: Perform dilation on a binary image.
  - *Erode*: Perform erosion on a binary image.
  - *Close (Dilate+Erode)*: Perform dilation followed by erosion on a binary image.
  - *Open (Erode+Dilate)*: Perform erosion followed by dilation on a binary image.

- NOTE: Image-Oriented methods may break covariance (add multivariate rank) because variable slabs handled separately
- Standard variable-space preprocessing can be used too, but are spatially insensitive

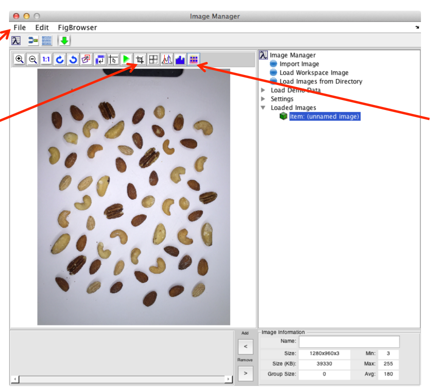
EIGENVECTOR RESEARCH INCORPORATED

Particle Analysis: Nuts

- Mixed nuts laid out on cutting board
- Photo taken with iPhone
- Under counter lighting plus flash
- In HW folder as Nuts3.jpg

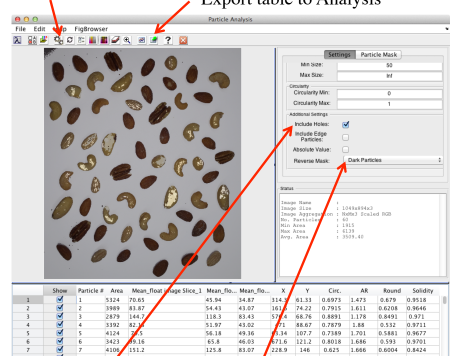
EIGENVECTOR RESEARCH INCORPORATED

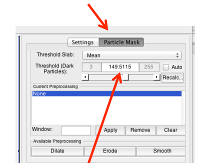
Import to Image Manager



EIGENVECTOR RESEARCH INCORPORATED

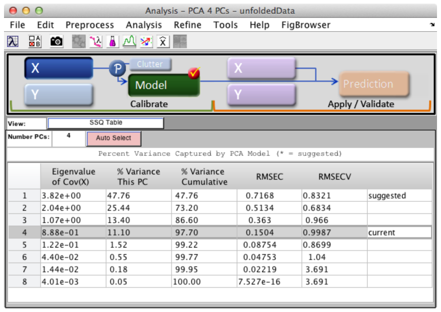
Particle Analysis Settings





EIGENVECTOR RESEARCH INCORPORATED

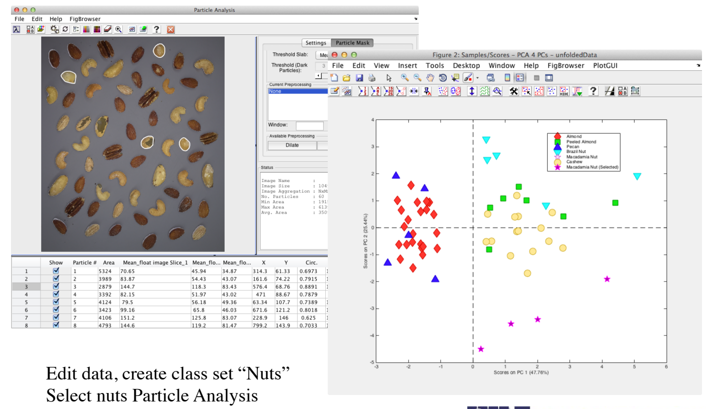
PCA on Particle Table



	Eigenvalue of Cov(X)	% Variance This PC	% Variance Cumulative	RMSEC	RMSECV	
1	3.82e+00	47.76	47.76	0.7168	0.8321	suggested
2	2.04e+00	25.44	73.20	0.5134	0.6834	
3	1.07e+00	13.40	86.60	0.363	0.966	
4	8.88e-01	11.10	97.70	0.1504	0.9987	current
5	1.22e-01	1.52	99.22	0.08754	0.8699	
6	4.40e-02	0.55	99.77	0.04753	1.04	
7	1.44e-02	0.18	99.95	0.02219	3.691	
8	4.01e-03	0.05	100.00	7.527e-16	3.691	

Autoscaled, X and Y variables omitted

Assign Classes to Particles



Edit data, create class set "Nuts"
Select nuts Particle Analysis
Assign classes

Further Possibilities and Improvements

- Would help to have better background subtraction
  - Use image of background with no particles and subtract
  - Fit function to background and substract
- Could build classifier (PLS-DA, etc.) based on particle statistics

Displaying a Multivariate Image
(4-10 Variables)

- How to choose the 3 variables?
  - In which order should they be displayed?
- Doesn't choosing ignore potential information in the remaining variables?
- How could information be extract from the image?
- What happens when we go to more variables? ...
- …. Factor-based techniques
  - use the correlation structure to enhance S/N
  - really good for hyperspectral

MIA: PCA-Based Methods

- Many methods are based on the spectroscopic information in an image
  - although spatial information is ignored mathematically
  - images are examined for spatial structure
- PCA (Principal Components Analysis)
  - Exploratory analysis
- SIMCA (Soft Independent Method Class Analogy)
  - Classification

EIGENVECTOR RESEARCH INCORPORATED

Image PCA

- Matricizing
- PCA: scores, scores images, loadings
  - unusual samples Q and T^2
  - score-score plots, density plots
  - linking scores and image plane(s)
  - contrast enhancement

EIGENVECTOR RESEARCH INCORPORATED

PCA Math Summary

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

$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \ldots + \mathbf{t}_K\mathbf{p}_K^T + \ldots + \mathbf{t}_R\mathbf{p}_R^T$$

  Where $R \leq \min\{M,N\}$, and the $\boldsymbol{t}_k\boldsymbol{p}_k^T$ pairs are ordered by the amount of variance captured.
- Generally, the model is truncated to *K* PCs, leaving some small amount of variance in a residual matrix **E**:

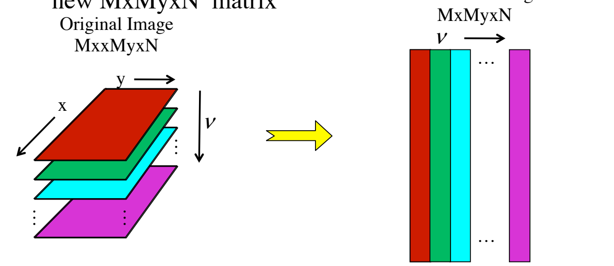
$$\mathbf{X} = \mathbf{t}_1\mathbf{p}_1^T + \mathbf{t}_2\mathbf{p}_2^T + \ldots + \mathbf{t}_K\mathbf{p}_K^T + \mathbf{E} = \mathbf{T}\mathbf{P}^T + \mathbf{E}$$

- where **T** is *M*x*K* and **P** is *N*x*K*.

EIGENVECTOR RESEARCH INCORPORATED

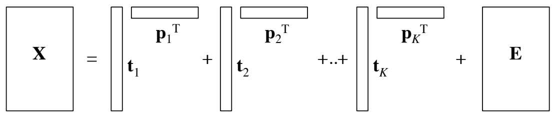
Matricizing (a.k.a. Unfolding)

- PCA works on X (MxN) but the image is MxxMyxN
  - reshape by matricizing such that each pixel is a row in a new MxMyxN matrix



EIGENVECTOR RESEARCH INCORPORATED

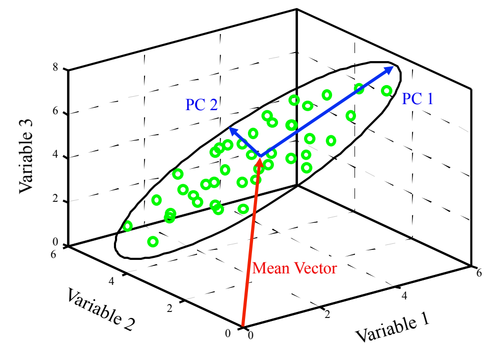
Properties of PCA



- $\mathbf{t}_k$,$\mathbf{p}_k$ ordered by amount of *variance captured*
  - λ_k are the eigenvalues of $\mathbf{X}^T\mathbf{X} \rightarrow \mathbf{X}^T\mathbf{X}\mathbf{p}_k = \lambda_k\mathbf{p}_k$
  - λ_k are $\propto$ variance captured
- $\mathbf{t}_k$ (*scores*) form an orthogonal set $\mathbf{T}_K$ (MxK)
  - describe relationship between *samples* → *pixels* ($M = M_x M_y$)
- $\mathbf{p}_k$ (*loadings*) form an orthonormal set $\mathbf{P}_K$ (NxK)
  - describe relationship between *variables*

EIGENVECTOR RESEARCH INCORPORATED

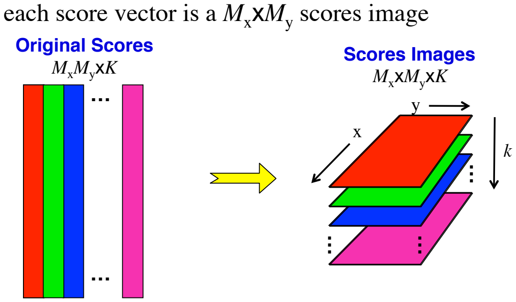
PCA Graphically



EIGENVECTOR RESEARCH INCORPORATED

Reshape Scores To Images

- PCA gives scores $\mathbf{T}$ (MxK) which is reshaped to scores images (M_xxM_yxK)
  - each score vector is a $M_x M_y$ scores image



EIGENVECTOR RESEARCH INCORPORATED

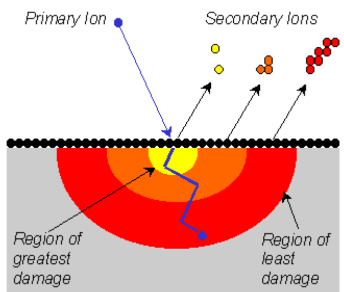
Plots / Images for PCA

- scores and loadings plots are interpreted in pairs
  - plot $\mathbf{t}_k$ vs sample number
    - find relationship between *samples* → *pixels*
    - each $M_x M_y$x1 score vector is reshaped to a M_xxM_y matrix that can be visualized as a "*scores image*" showing spatial relationships between pixels
  - $\mathbf{p}_k$ vs variable number
    - relationship between *variables* responsible for observations in samples
- it is useful to plot $\mathbf{t}_{k+1}$ vs. $\mathbf{t}_k$ and $\mathbf{p}_{k+1}$ vs. $\mathbf{p}_k$
  - examine image and score / score plots

EIGENVECTOR RESEARCH INCORPORATED

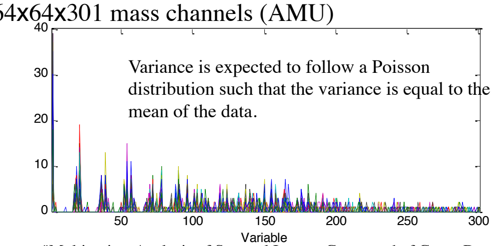
TOF-SIMS of PMMA and Deuterated Polystyrene

- Time of flight secondary ion mass spectroscopy used for surface analysis
- Mass spectrum for each pixel
- Thanks to Physical Electronics for the data

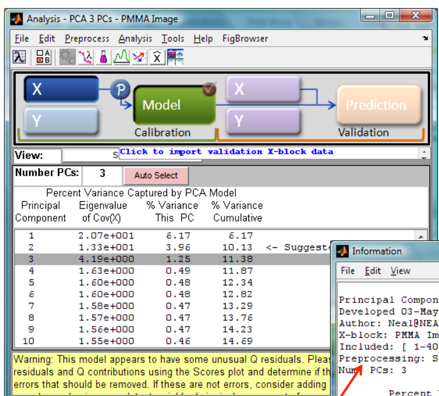


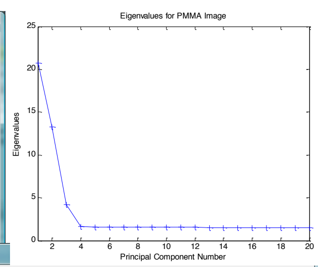
Example Data

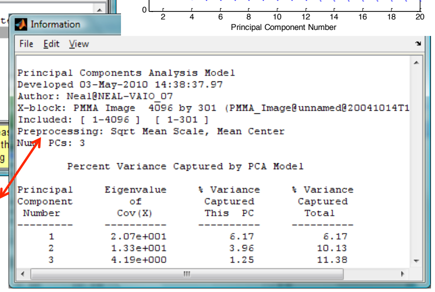
- Data is positive SIMS spectrum at each pixel (point) on a 64x64 grid
- 64x64x301 mass channels (AMU)



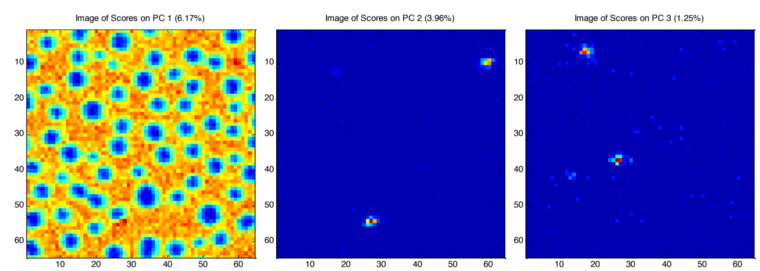
M.R. Keenan, "Multivariate Analysis of Spectral Images Composed of Count Data," in *Techniques and Applications of Hyperspectral Image Analysis*, H. F. Grahn and P. Geladi, eds. (John Wiley & Sons, West Sussex, England), 89-126, 2007.







Poisson scaling, and mean-centering



Scores images show islands of polystyrene in PMMA and two sources of unusual variance

PCA Statistics

- Limits can be set for
  - Q residual: lack of fit statistic
    - for a row of $\mathbf{E}$, $\mathbf{e}_m$, and a row of $\mathbf{X}$, $\mathbf{x}_m$, $m = 1, \ldots, M$

$$Q_m = \mathbf{e}_m \mathbf{e}_m^T = \mathbf{x}_m (\mathbf{I} - \mathbf{P}_K \mathbf{P}_K^T) \mathbf{x}_m^T$$

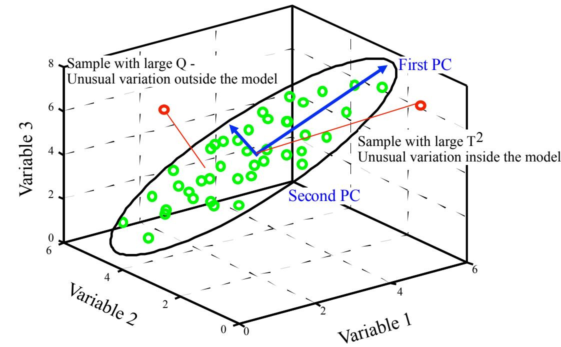
  - Hotelling's T^2 statistic
    - for a row of $\mathbf{T}_K$, $\mathbf{t}_m$, and KxK diagonal matrix λ

$$T_m^2 = \mathbf{t}_m \lambda^{-1} \mathbf{t}_m^T = \mathbf{x}_m \mathbf{P}_K \lambda^{-1} \mathbf{P}_K^T \mathbf{x}_m^T$$

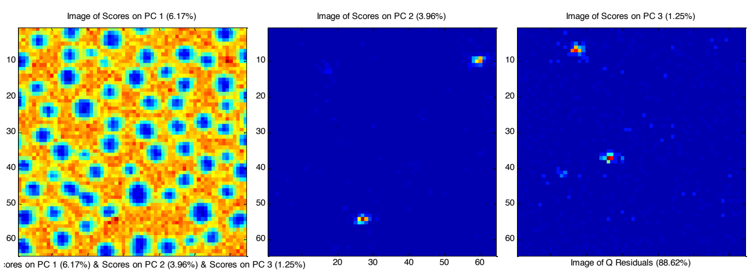
  - and also for individual columns:
    - scores, $\mathbf{t}_{mk}$
    - residuals $\mathbf{e}_{mk}$

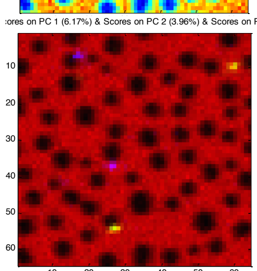
EIGENVECTOR RESEARCH INCORPORATED

Geometry of Q and T^2



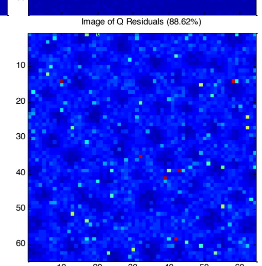
EIGENVECTOR RESEARCH INCORPORATED



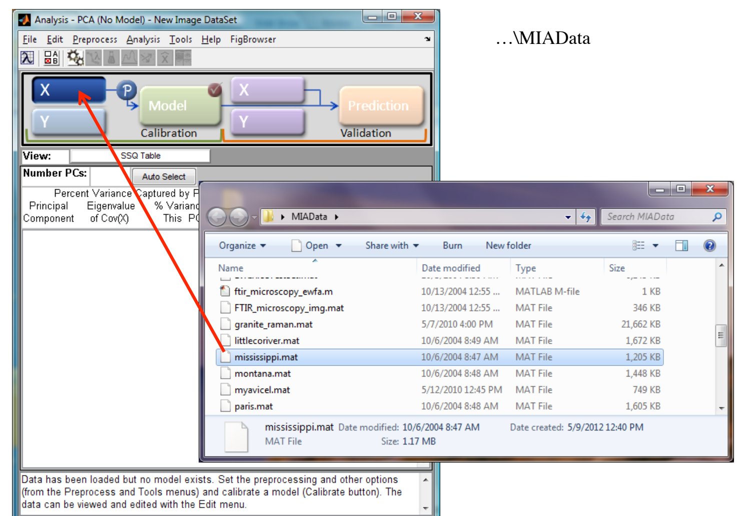


RGB of scores on PCs 1, 2 and 3

Q residuals

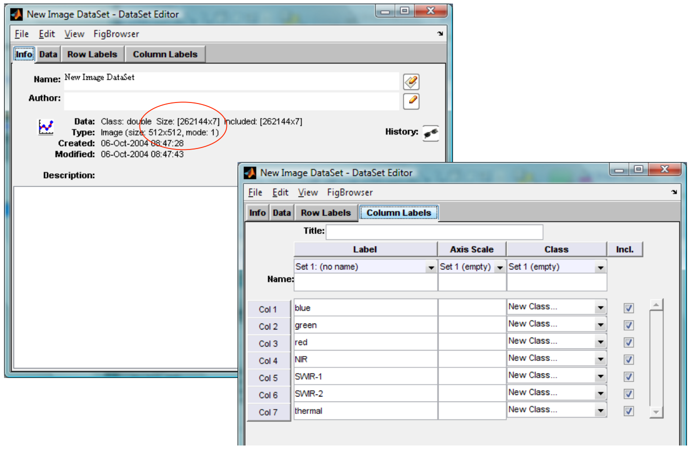


EIGENVECTOR RESEARCH INCORPORATED



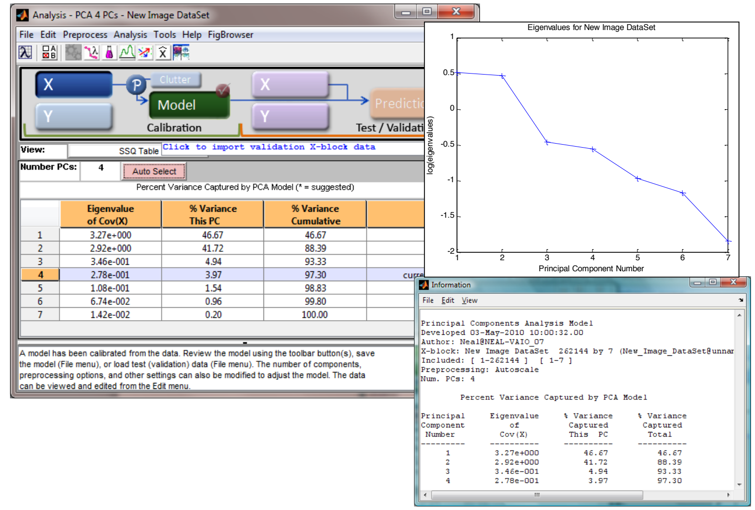
…\MIAData

EIGENVECTOR RESEARCH INCORPORATED



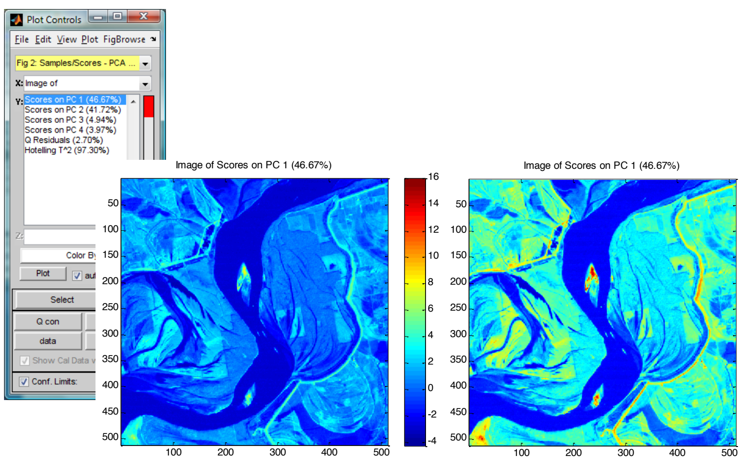
76

EIGENVECTOR RESEARCH INCORPORATED



77

EIGENVECTOR RESEARCH INCORPORATED

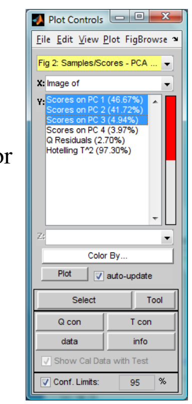


78

EIGENVECTOR RESEARCH INCORPORATED

Creating Color Images

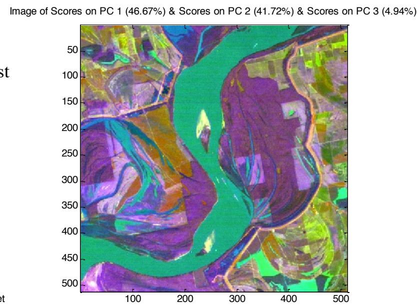
- Images are made of three colors: red, green and blue
  - e.g., scaled to integers 0-255 for 8-bit color
- Scores can be used to define the colors
  - PC 1 = red, PC 2 = green, PC 3 = blue



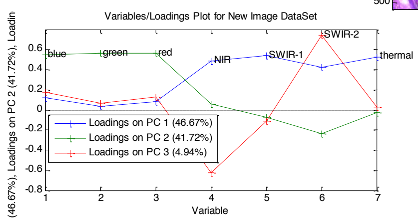
79

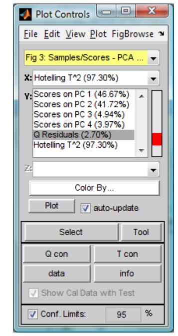
EIGENVECTOR RESEARCH INCORPORATED

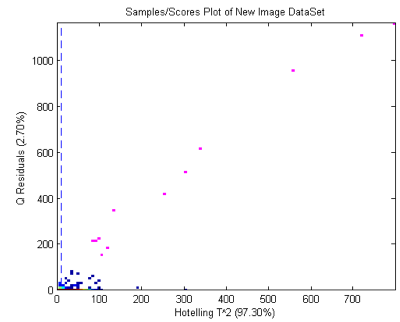
Scores
View:Auto Contrast



Loadings
View:Labels:Set 1



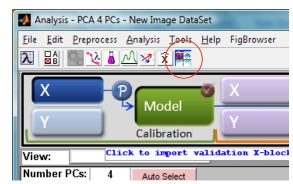


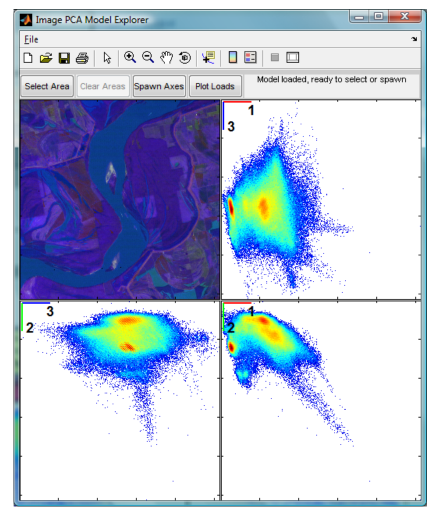


pixels with high Q and T2 have been selected

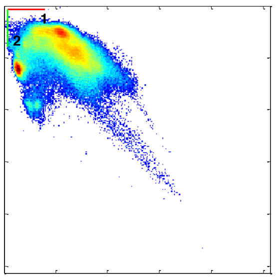
Bivariate Scores Plots

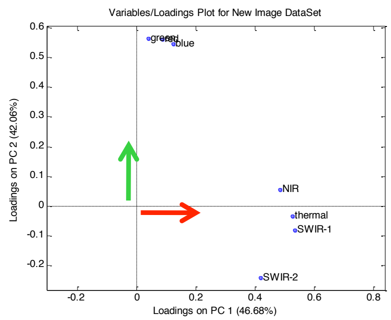
- Plotting $\mathbf{t}_{k+1}$ vs. $\mathbf{t}_k$ (score / score plots)
- Problem: lot's of points
  - 512*512 = 262144 points with lot's of them falling on top of each other (big blobs)
- Density plots
  - count the number of points that lie on top of each other (have same score / score value)
  - color code according to density
  - use log to allow easy comparison between large and small number densities



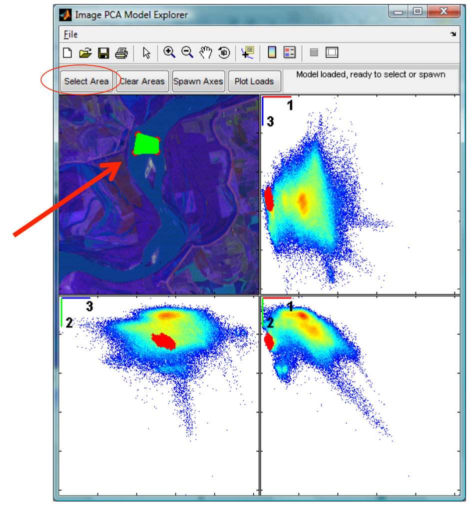


Scores and Loadings



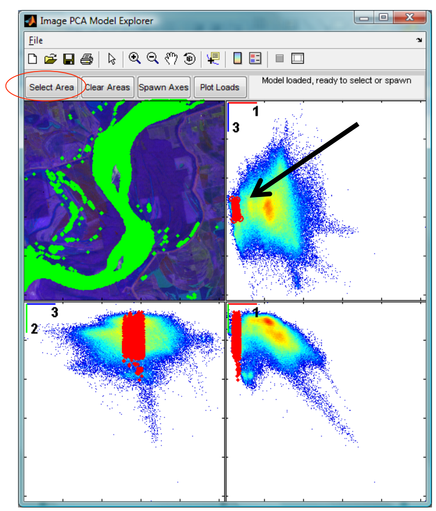


EIGENVECTOR RESEARCH INCORPORATED



selecting an area w/in the image plane shows where it lies in the scores space

EIGENVECTOR RESEARCH INCORPORATED



selecting an area w/in the scores space shows where it lies in the image plane

images can be explored to find similarities and differences w/in an image

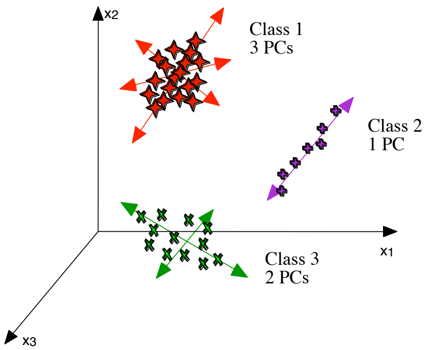
EIGENVECTOR RESEARCH INCORPORATED

SIMCA

- Supervised pattern recognition / classification technique
  - the model is a collection of PCA models
  - each "class" is a separate PCA model
  - new samples are compared to all of the PCA models and scores, T^2 and Q are compared to statistical limits on each model
  - samples can belong to one, none or more than one class

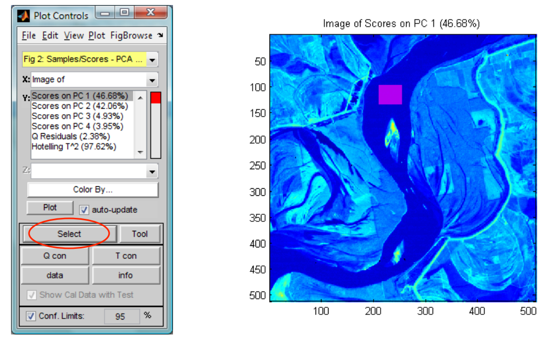
EIGENVECTOR RESEARCH INCORPORATED

A SIMCA Model



SIMCA Example

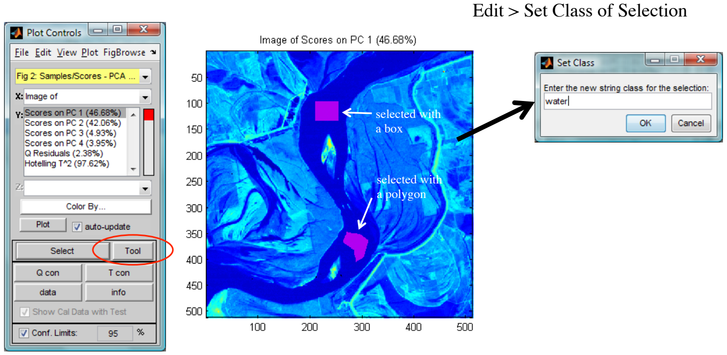
For SIMCA, classes need to be defined.
Use the selection tool to select regions in the image that are expected to be similar and to be modeled as a single class.



SIMCA Example

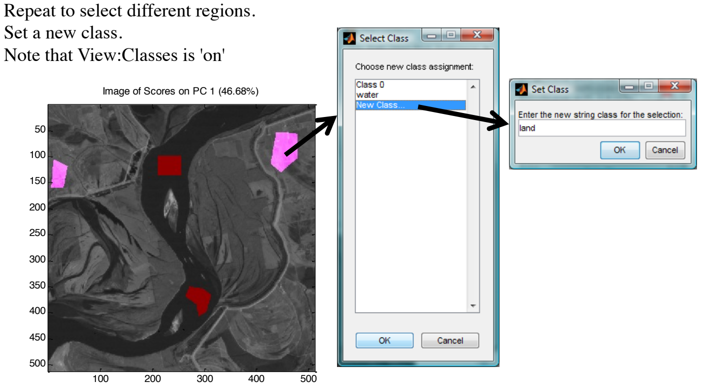
- Use the Tool to change the selection tool.
- Hold shift to select multiple regions.

Edit > Set Class of Selection



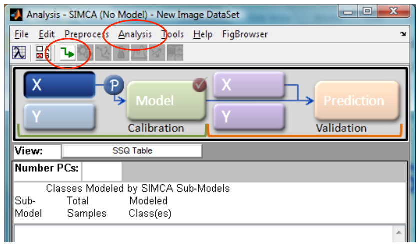
SIMCA Example

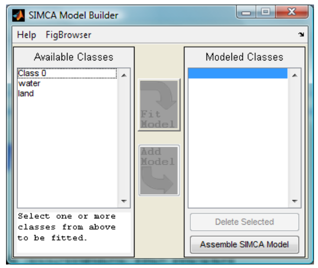
- Repeat to select different regions.
- Set a new class.
- Note that View:Classes is 'on'



SIMCA Model Builder

- SIMCA requires a selection of classes to be modeled and then assembles the model
  - Analysis:SIMCA

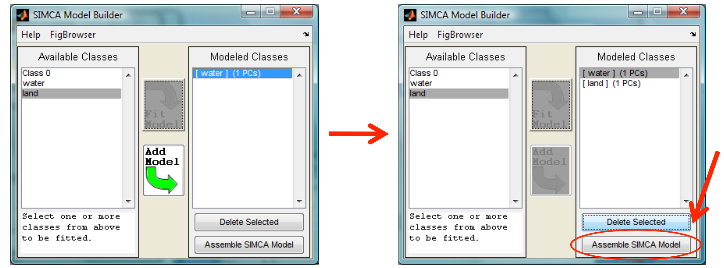




EIGENVECTOR RESEARCH INCORPORATED

Model of Each Class

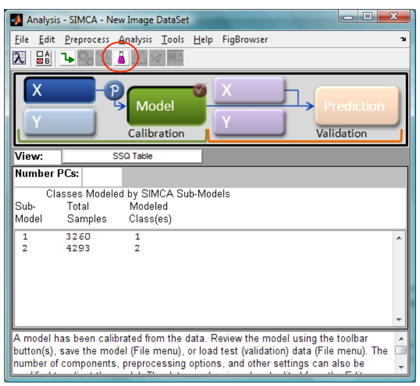
- Each class is modeled using PCA
  - highlight a class and then "fit model"
  - select the number of PCs, etc., then "add model"



EIGENVECTOR RESEARCH INCORPORATED

SIMCA Example

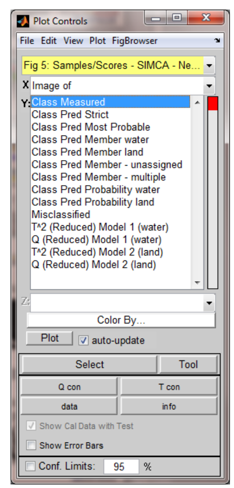
- The SIMCA model consists of two PCA models



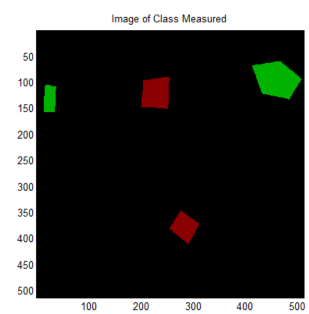
- Data from the entire image will be projected onto each PCA model.
- Scores, Q and T^2 are calculated for each model and it is determined which model the data is closest to.
- Click the scores button to examine the images.

EIGENVECTOR RESEARCH INCORPORATED

SIMCA Model Predictions



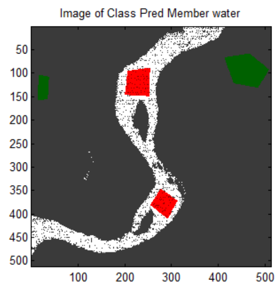
- "Class Measured" = where the classes were selected.
- "Reduced" means that the statistic was normalized by the limit of the corresponding statistic (e.g., to the 95% CL).

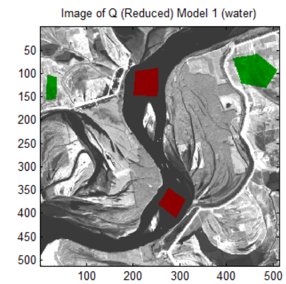


EIGENVECTOR RESEARCH INCORPORATED

Model 1 Predictions

- Model 1 (w/in set limits for both Q and T^2)
- Reduced Q on Model 1 (dark is low)

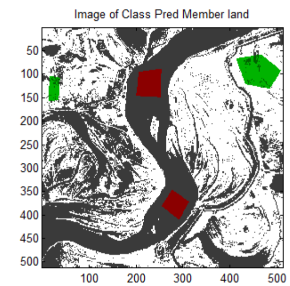


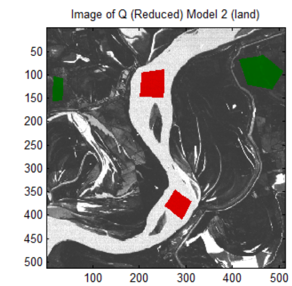


EIGENVECTOR RESEARCH INCORPORATED

Model 2 Predictions

- Model 2 (w/in set limits for both Q and T^2)
- Reduced Q on Model 2 (dark is low)

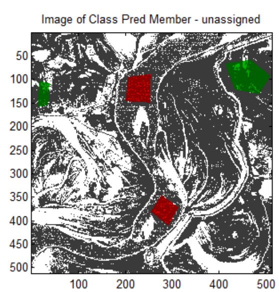


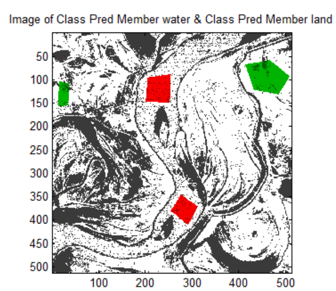


EIGENVECTOR RESEARCH INCORPORATED

In Model and Not-In-Any Model

- Outside of both models (left)
- Inside either model (right)

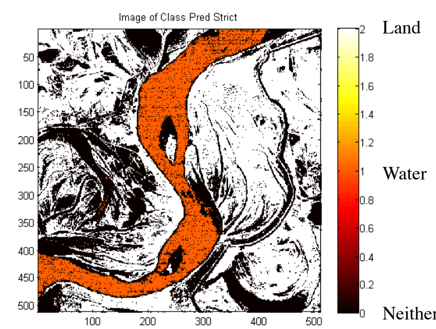




EIGENVECTOR RESEARCH INCORPORATED

"Strict" Class Predictions

- Strict predictions require probability of 50% or greater for one class only
- (Note: turn off classes to view)



EIGENVECTOR RESEARCH INCORPORATED

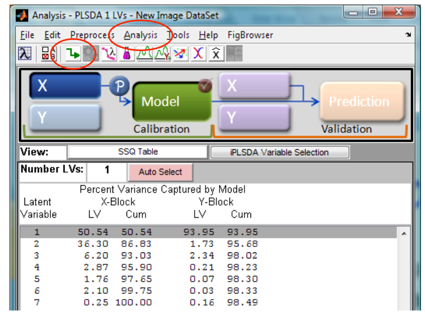
Image PCA Conclusions

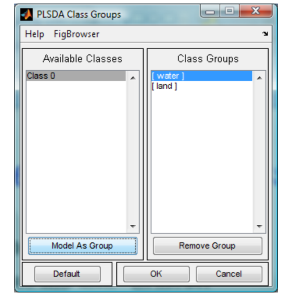
- Image PCA is a useful unsupervised pattern recognition technique for exploring images
  - scores and loadings are useful for determining what original variables are responsible for differences observed in an image
    - score-score plots and linked score plots
    - contrast enhancement might be needed to see small changes
- Image SIMCA is a useful supervised pattern recognition technique
  - find similar / dissimilar portions of an image very quickly

EIGENVECTOR RESEARCH INCORPORATED

PLSDA Model Builder

- PLS discriminant analysis requires a selection of classes to be modeled
  - Analysis:PLSDA

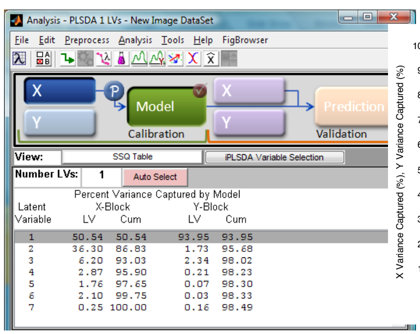


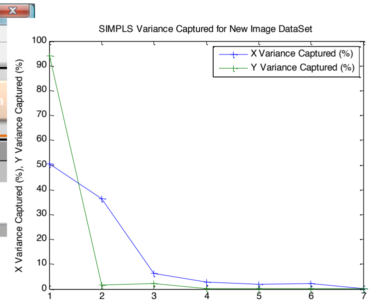


EIGENVECTOR RESEARCH INCORPORATED

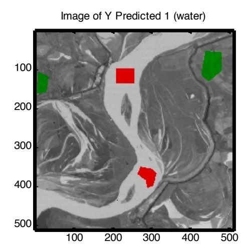
PLSDA Maximizes Class Separation on a PLS Model

- PLS (selection of factors, cross-validation, etc.)

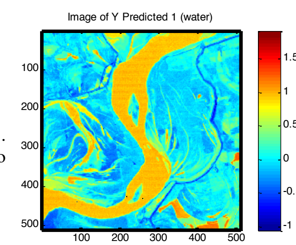


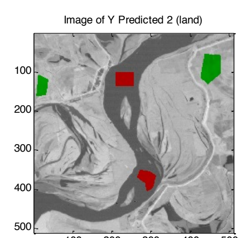


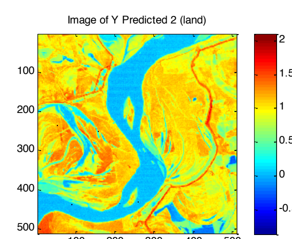
EIGENVECTOR RESEARCH INCORPORATED



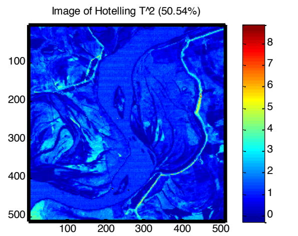
- Data from the entire image are projected onto the PLSDA model.
- Light shows high predictions on each class.
- Click the scores button to examine the images.
- View:Classes (uncheck Set 1)

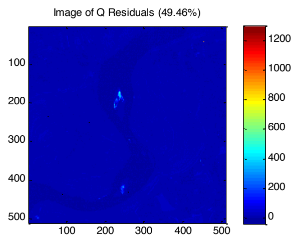




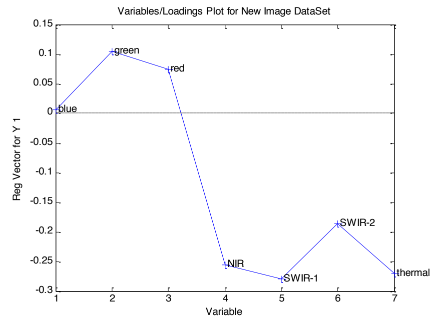


EIGENVECTOR RESEARCH INCORPORATED



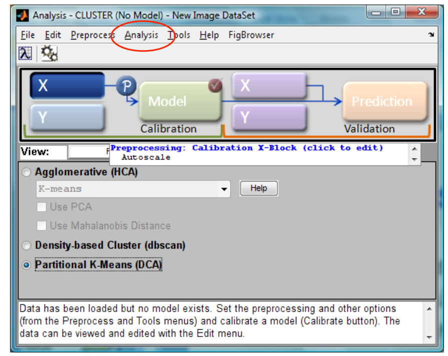


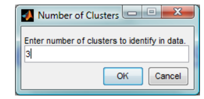
- Inspect T^2 and Q
- Regression vector suggests that green and red increase relative to IR channels for water relative to land



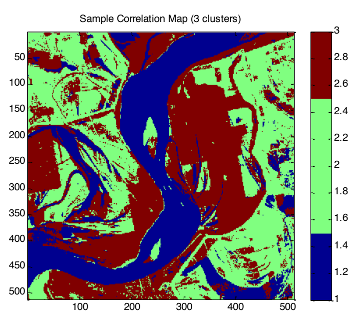
Cluster Analysis

- Analysis:Cluster





Results for 3 and 5 Clusters



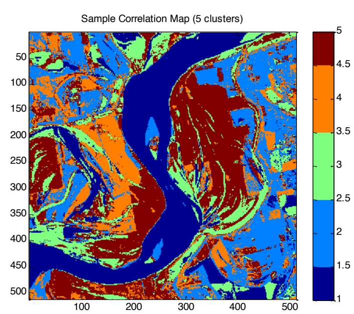


Image PLSDA and Clustering Conclusions

- If classes (regions) are known, PLSDA is a useful supervised pattern recognition technique for exploring images
  - can often bring out more contrast than PCA
- Image clustering is a useful unsupervised pattern recognition technique (guess number of clusters)
  - find similar / dissimilar portions of an image very quickly
- Results of all analysis methods must be consistent

Comments on Presenting Images

- Images are representations of spatial and chemical information, …
- but they can be mis-used.
  - users can control colors and contrasting and select channels or PCs (or rotations thereof)
  - as a result some things can be highlighted while others can be hidden
- It is important to report how images were constructed
  - the work must be reproducible

EIGENVECTOR RESEARCH INCORPORATED

MCR Objective

- With a minimum of *a priori* information, decompose a data matrix into chemically meaningful factors
  - "pure analyte" spectra (in contrast to loadings and weights)
  - "pure analyte" concentrations (in contrast to scores)
- Easy to interpret
  - can be used for process monitoring, QC, ...
  - improving model performance (e.g., regression)
    - can include constraints on predicted concentrations for greater user control

Anna de Juan and Romà Tauler, "Multivariate Curve Resolution (MCR) from 2000: Progress in Concepts and Applications," *Crit. Rev. Anal. Chem.*, **36**:163-176 2006.

EIGENVECTOR RESEARCH INCORPORATED

Curve Resolution – Motivation

Use observed **correlations among samples and variables** from **multiple measurements** to determine:

- The **number** of components in the system
  - using knowledge of chemistry and physics and also
  - rank analysis with PCA and evolving factor analysis
- The chemical / physical **characteristics** of the components / factors (e.g., spectral shape)
- The **quantity** of the components in each sample
- Also need to know when the objective cannot be met and the source of potential ambguities

EIGENVECTOR RESEARCH INCORPORATED

Classical Least Squares

- Classical Least Squares (CLS)
  - commonly used with spectra

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

- Useful for estimating **C** when *all* *K* analyte spectra are known
  - $\mathbf{X}_{M\mathrm{x}N}$ are measured spectra
    - **X** can be an "unfolded" image where *M* is the total number of pixels and *N* is the number of channels
  - $\mathbf{C}_{M\mathrm{x}K}$ are concentrations
  - $\mathbf{S}_{N\mathrm{x}K}$ are pure analyte spectra

EIGENVECTOR RESEARCH INCORPORATED

Classical Least Squares for C

- Given **S** (spectra), the **C** (concentrations) are found by minimizing:

  $$\mathbf{E}\mathbf{E}^T = (\mathbf{X}\text{-}\mathbf{C}\mathbf{S}^T)(\mathbf{X}\text{-}\mathbf{C}\mathbf{S}^T)^T$$

  with respect to **C** resulting in

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

```
>> c = x/s;
```

- Can also use non-negativity constraints (i.e., negative concentrations are not allowed)

```
>> c = fasternnls(s',x')';
```


EIGENVECTOR RESEARCH INCORPORATED

Classical Least Squares for S

- Given **C** (concentrations), the **S** (spectra) are found by minimizing:

  $$\mathbf{E}^T\mathbf{E} = (\mathbf{X}\text{-}\mathbf{C}\mathbf{S}^T)^T(\mathbf{X}\text{-}\mathbf{C}\mathbf{S}^T)$$

  with respect to **S** resulting in

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

```
>> s = c\x;
```

- Can also use non-negativity constraints (i.e., negative intensities are not allowed)

```
>> s = fasternnls(c,x);
```


EIGENVECTOR RESEARCH INCORPORATED

Alternating Least Squares (ALS)

- What if we don't know **S** or **C**?
- Given *initial guess* $\mathbf{S}_0$ (or $\mathbf{C}_0$)...

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

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

- Iterate until convergence
  - Usually non-negatively constrained (**C**>0 and **S**>0)
  - and each $\mathbf{s}_k^T\mathbf{s}_k=1$ (i.e., unit length **S** vectors)
- Most popular method for multivariate curve resolution (MCR)

  a.k.a. self-modeling curve resolution, self-modeling mixture analysis, end-member extraction

EIGENVECTOR RESEARCH INCORPORATED

MCR Model Estimation

Geometrical Approaches

- purity, SIMPLSMA, DISTSLCT
- simple, fast
  - useful for quick qualitative interpretation
- can find small factors
  - useful for outlier detection
  - but adversely affected by outliers
- doesn't typically apply constraints
  - selectivity is necessary for a good soln
  - some solutions not physically meaningful and the spectral basis may not be useful for application to future data
- minimize the Frobenius norm (may not iterate to the final solution)
- often used as a good first guess for least squares approaches

Least Squares

- constrained alternating least squares, positive matrix factorization
- mathematically rigorous, slow
- small factors can be "lost in the variance"
  - less affected by outliers
- applies physically meaningful constraints
  - can be difficult to id proper constraints
  - basis often useful for future application
  - can be modified to be quantitative
- typically minimizes the Frobenius norm subject to constraints

EIGENVECTOR RESEARCH INCORPORATED

Preprocessing for MCR

- Although mean centering is almost always done in other multivariate methods, it is almost *never* done in MCR.
  - Zero has importance in many MCR models (as it does with CLS)
  - Non-negativity constraint *can not* be used if mean-centering is done (by definition, some data is below zero after mean-centering so some spectra or concentrations would have to be negative!)
  - An offset (e.g. baseline) can be fit as a separate factor.
  - mean-centering may be useful for non-spectra data
- Because many MCR methods utilize least-squares step(s), adjustment of scales may be critical.
  - Normalization of samples (e.g., normalize, SNV, MSC)
  - Normalization of variables (e.g., Poisson (sqrt mean scale), autoscale)

EIGENVECTOR RESEARCH INCORPORATED

Preprocessing for MCR

- Overall Goal: Remove distraction
  - For example, exclude "bad" variables/samples.
    - usually perform an exploratory analysis prior to attempting MCR
- Sample normalization

  Questions:
  - Does response scale matter to discrimination?
  - Are there other interferences which may affect normalization?
- Other Preprocessing
  - Derivatives? Not with non-negativity, constraint must be relaxed.
    - Makes interpretation harder but might provide more interpretable results in other mode.
  - Background subtraction, baselining? ← May be questionable.
    - Can use fixed component and allow least-squares to solve subtraction (ala Extended Least Squares – see also constraints later in this section)

EIGENVECTOR RESEARCH INCORPORATED

Rotational Ambiguity

- For an invertible, non-diagonal square matrix **A** all solutions have the same fitness **E**:

$$\mathbf{X} = \mathbf{C}\mathbf{S}^T + \mathbf{E} = (\mathbf{C}\mathbf{A})(\mathbf{A}^{-1}\mathbf{S}^T) + \mathbf{E}$$

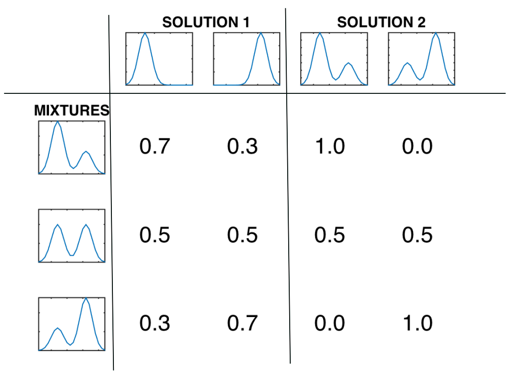
  this is a *rotational ambiguity*
- Two solutions with equivalent fit (error)

$$\mathbf{C}_a\mathbf{S}_a^T + \mathbf{E} = \mathbf{C}_b\mathbf{S}_b^T + \mathbf{E} \quad where: S_a \neq S_b$$

Tauler, R., "Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution", *J. Chemo.*, **15**, 627-646, 2001.

EIGENVECTOR RESEARCH INCORPORATED

Rotational Ambiguity



EIGENVECTOR RESEARCH INCORPORATED

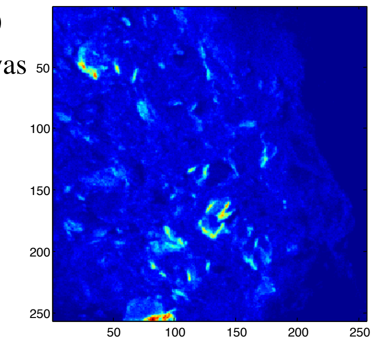
How Many Components, K?

- ALS-based methods can be quite sensitive to the number of components K selected.
- Too many will cause degeneracy of a component
  - degeneracy is when one factor splits into two or more factors
- Good estimate for K comes from a conservative PCA model estimate.
  - Look for significance of eigenvalues and structure in loadings and scores.
- Slowly increase the number components and evaluate *all* recovered components in the model.

EIGENVECTOR RESEARCH INCORPORATED

Imaging Mass Spec

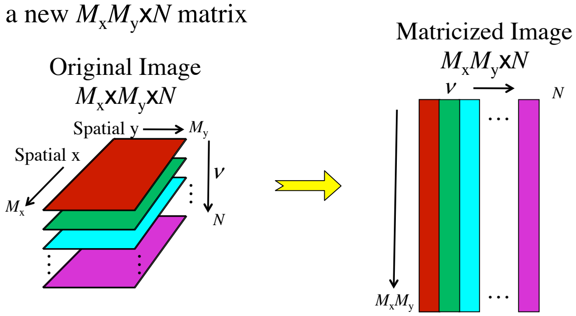
- Image is 256x256x90
- The mass spectrum was 41945 mass channels selected and binned into 93 channels
- Image of total ion count
  - false color



EIGENVECTOR RESEARCH INCORPORATED

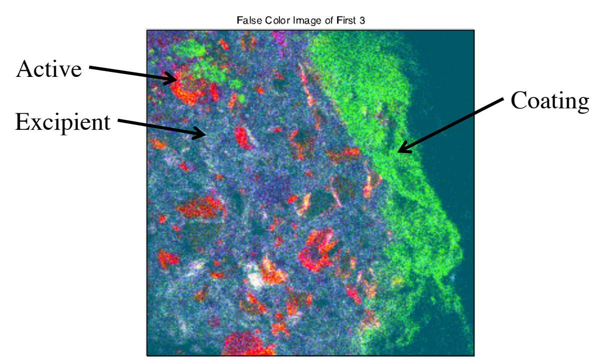
MCR on Images (via Unfolding)

- MCR works on **X** (MxN) but the image is M_xxM_yxN
- Reshape by "matricizing" such that each pixel is a row in a new M_xM_yxN matrix



EIGENVECTOR RESEARCH INCORPORATED

PCA Score Image

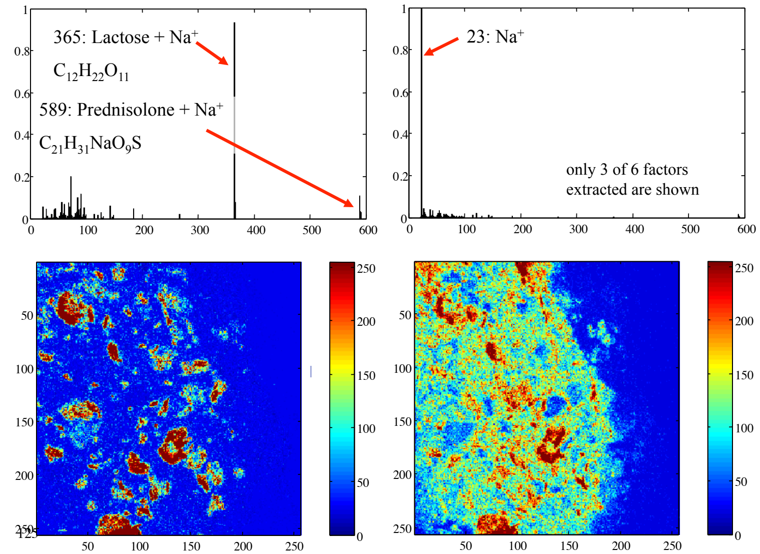


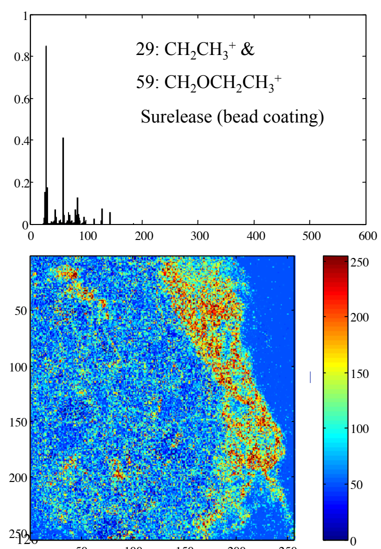
EIGENVECTOR RESEARCH INCORPORATED

MCR (ALS) on TOF-SIMS Image

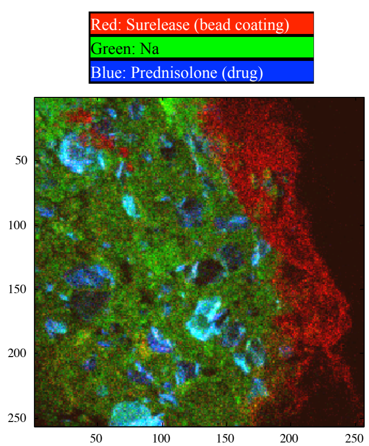
- Non-negative constraints on both **C** and **S**
- Initialize with pure/extreme samples (i.e., pixels)
- Recover 6 interpretable spectra and concentration profiles (matricized “scores” or “contributions”)
- Show score Images
  - image was matricized for MCR decomposition
  - scores are rearranged back to form contribution images
    - result is chemical imaging

Gallagher, N.B., Shaver, J.M., Martin, E.B., Morris, J., Wise, B.M. and Windig, W., “Curve resolution for images with applications to TOF-SIMS and Raman”, *Chemometr. Intell. Lab.*, **73**(1), 105–117 (2003).



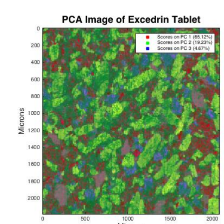


RGB “Chemical” Image



Example: MCR on Excedrin

- Excedrin is a mixture of aspirin, acetaminophen, caffeine and microcrystalline cellulose
- Tablet imaged with tunable laser from 800 to 1800 cm^{-1} over ~2mm
- Thanks to Agilent for data!

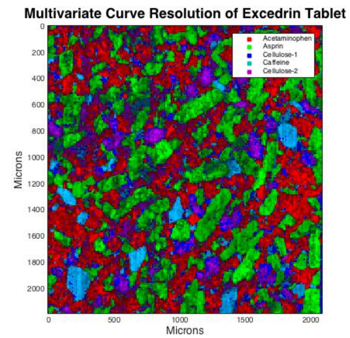


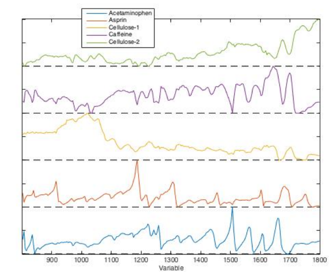
Perform MCR on Excedrin

- Load Excedrin_sm into Analysis in MCR mode
- Set preprocessing to "none"
- Set number of components to 5
- Select "auto contrast" for score images
- Chemical species can be assigned to components based on known features

```
legend('Acetaminophen','Asprin','Cellulose-1','Caffeine','Cellulose-2')
```


MCR on Excedrin Results





Further possibilites

- Export score images to particle analysis
  - Determine particle size distributions of ingredients
  - Check formulation for composition
- Convert MCR model to CLS model
  - Extract loadings from MCR model
  - Load as CLS model
  - Assign component names
  - Use on new images

Other Ways of Focusing on Variance of Interest

- Maximum Autocorrelation Factors – find variance with spatial correlation
- Maximum Difference Factors – find variance with spatial transitions (multivariate edge detection)
- Generalized Least Squares Weighting – ignore variance from specified regions

Maximum Autocorrelation Factors for Multivariate Images

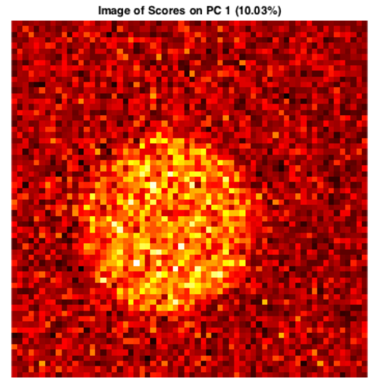
- For MAF, the clutter is the first spatial difference
  - the first difference should be high on edges and just noise w/in clusters
- For MNF, the clutter is intra-class variance
  - the result is the same generalized eigenvector problem as MAF with different clutter $\mathbf{\Sigma}_C$

T.A. Blake, J.F. Kelly, N.B. Gallagher, P.L. Gassman and T.J. Johnson, "Passive detection of solid explosives in Mid-IR hyperspectral images," *Anal Bioanal Chem*, **395**, 337-348, 2009.

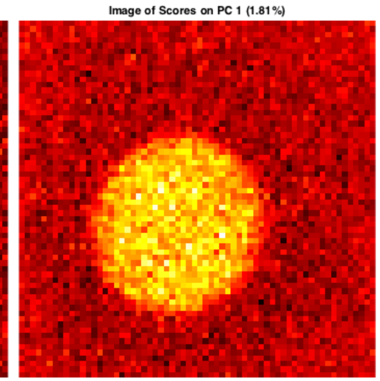
N.B. Gallagher, J.F. Kelly, T.A. Blake, "Passive infrared hyperspectral imaging for standoff detection of tetryl explosive residue on a steel surface," Whispers 2010, June 14-16, Reykjavik, Iceland

EIGENVECTOR RESEARCH INCORPORATED

MAF on SIMS Image of PVA



PCA



MAF

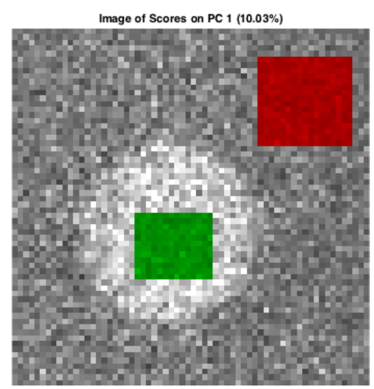
EIGENVECTOR RESEARCH INCORPORATED

Clutter Filters

- Define areas where only variance is due to noise or other unwanted variation
- Develop filter to minimize this variance
  - Generalized Least Squares (GLS) Weighting
    - Inverse square root of clutter covariance
  - External Parameter Orthogonalization (EPO)
    - Project out first PCs of clutter covariance

EIGENVECTOR RESEARCH INCORPORATED

Define Clutter Areas



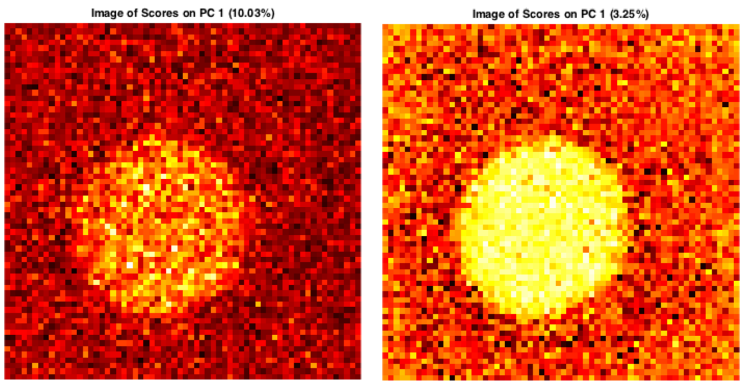
Only variation in marked areas is due to “noise”

Center each area to its own mean, then combine areas

Develop GLS weighting from combined areas

EIGENVECTOR RESEARCH INCORPORATED

GLS Filtered PVA



PCA

PCA with GLS

EIGENVECTOR RESEARCH INCORPORATED

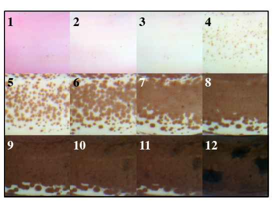
Mulitvariate Image Regression

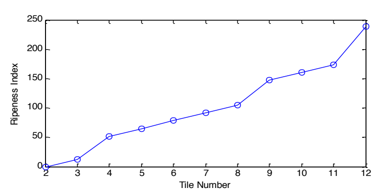
- Inverse least squares models
  - PCR, PLS
  - Similar to PCA for X-block
    - matricizing, scores, scores images, loadings, unusual samples Q and T^2, score-score plots, density plots, linking scores and image plane(s), contrast enhancement
  - Add predictions of a y-block
    - $\mathbf{y = Xb}$
    - predict a property
    - used for PLS-descriminant analysis

EIGENVECTOR RESEARCH INCORPORATED

Banana Ripeness by PLS

- Goal: Develop an automated (objective) method to assess banana ripeness
- X-Block RGB Images of Bananas at various stages of ripeness (Tiled)
- Y-Block Ripeness index for each tile

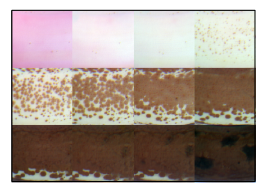




Data Courtesy Kim Esbensen
University of Ålborg, Denmark

EIGENVECTOR RESEARCH INCORPORATED

Two-Dimensional Calibration Data



X-block

Y-block

Image-based calibration takes advantage of high sampling rate of imaging (40 thousand samples for each tile!)

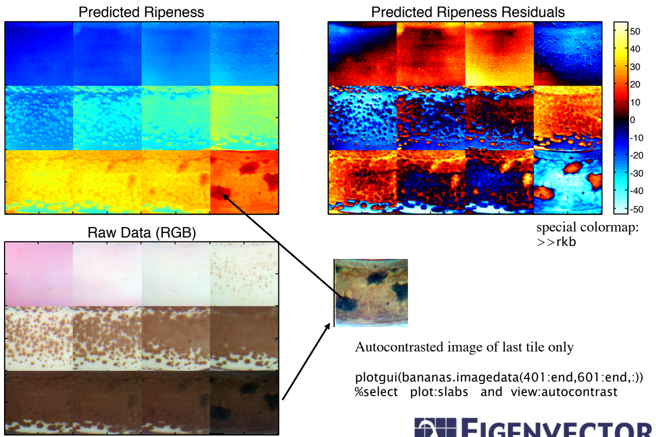
Y-block assumes a constant reference value for each image.

Unfold blocks before PLS

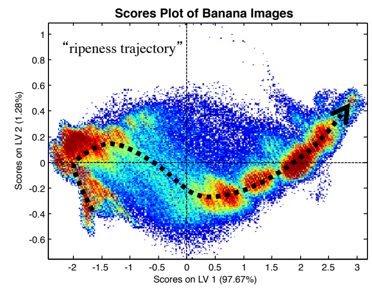
Note: Does not inherently take spatial correlation into account.

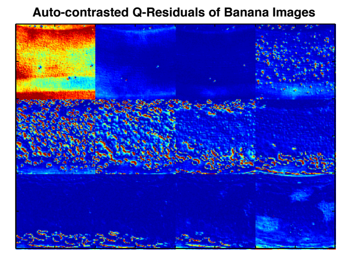
EIGENVECTOR RESEARCH INCORPORATED

Banana Predictions



Banana Scores and Q Residuals





Conclusions

- Anything that can be done with 2-way data tables can also be done with images
- Plus many other tools, e.g. particle analysis
- Special tools available to take advantage of spatial correlation
- Visually appealing!