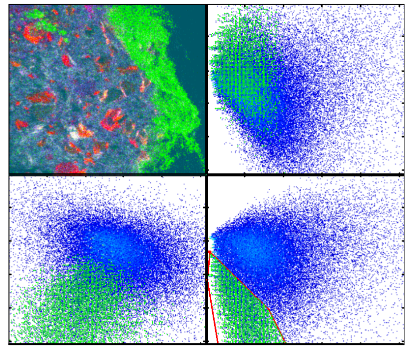


Introduction to Multivariate Image Analysis (MIA)

EIGENVECTOR RESEARCH INCORPORATED

Table of Contents

- Intro to 3-way arrays and simple visualizations and size/ shape analyses
- Practical Multivariate Image Analysis (MIA)
  - PCA, SIMCA, PLSDA and clustering
- Variance Filtering for Images:
  - Maximum Autocorrelation Factors, Maximum Difference Factors, Generalized Least Squares Weighting (MAF, MDF, GLSW)
- Multivariate Image Regression and Quantiative Analyses
  - Partial Least Squares, Classical Least Squares and Multivariate Curve Resolution Models (PLS, CLS, MLR)

EIGENVECTOR RESEARCH INCORPORATED

Resources

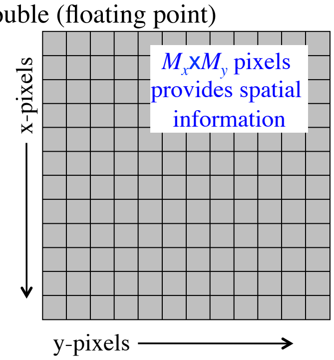
- *Hyperspectral Image Analysis*, eds. P. Geladi and H. Grahn, Wiley (2007), ISBN 978-0-470-01086-0
- *Chemometrics*, M.A. Sharaf, D.L. Illman and B.R. Kowalski, Wiley-Interscience (1986) ISBN 0-471-83106-9
- *Multivariate Analysis*, K.V. Mardia, J.I. Kent and J.M. Bibby, Academic Press, (1979) ISBN 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

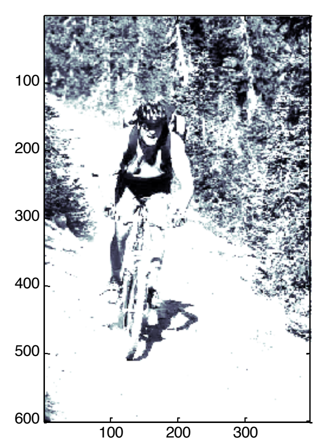
- Journal of Chemometrics
- IEEE Trans. on Geosci. and Remote Sensing
- Chemometrics and Intelligent Laboratory Systems
- Analytical Chemistry
- Analytica Chemica Acta
- Applied Spectroscopy
- Critical Reviews in Analytical Chemistry
- Journal of Process Control
- Computers in Chemical Engineering
- Technometrics
- ....

EIGENVECTOR RESEARCH INCORPORATED

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)



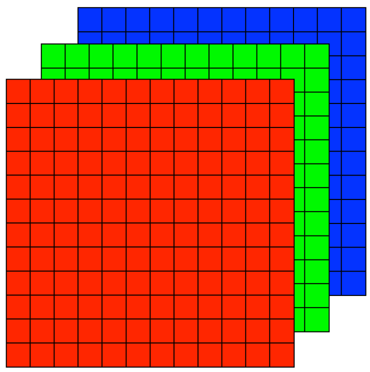


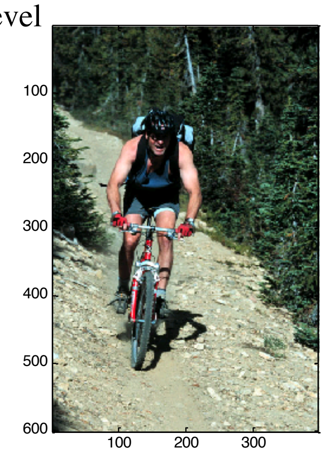
EIGENVECTOR RESEARCH INCORPORATED

Multivariate Image

(3 Variables)

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





EIGENVECTOR RESEARCH INCORPORATED

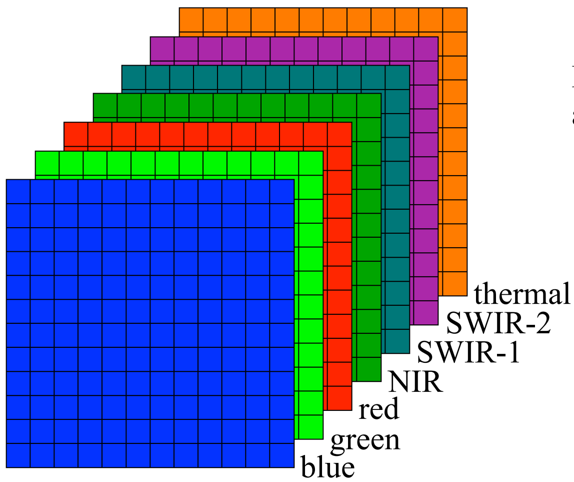
Image Analysis

- 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

EIGENVECTOR RESEARCH INCORPORATED

Multivariate Image
(4-10 Variables)

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

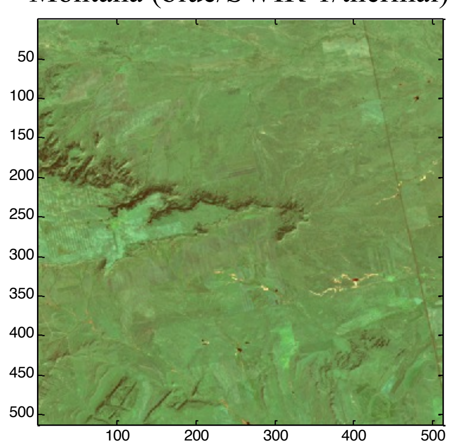


How should we display a seven variable image?

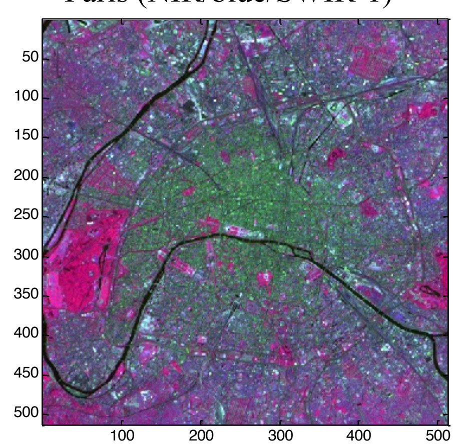
Multivariate Image
(4-10 Variables)

- Choose 3 of 7 (Landstat)

Montana (blue/SWIR-1/thermal)



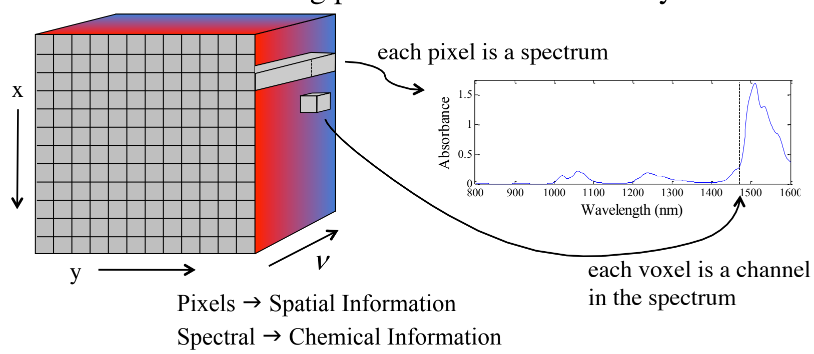
Paris (NIR/blue/SWIR-1)*



*contrast enhanced

Hyperspectral Image
(>10 Variables)

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

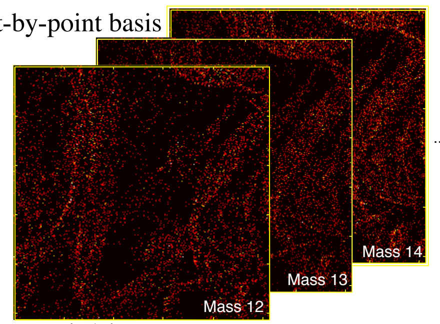


Multivariate Images

- 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, …

Physics of Measurement

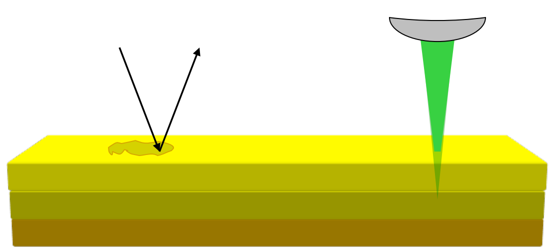
- Point scanning
  - spectra measured on a point-by-point basis
  - secondary-ion mass spec
  - atomic force microscopy
  - surface Raman
- Line scanning
  - push broom
- Focal plane array
  - images can be acquired very quickly



Volumetric Analysis Techniques

- Confocal Wavelength Resolved Imaging
- Surface Ablation Techniques

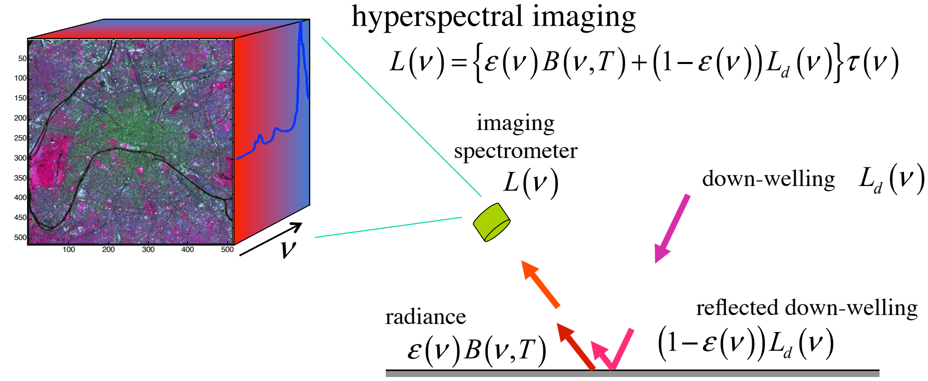
- Produces multivariate data in *3-dimensional space*



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

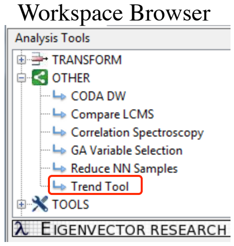
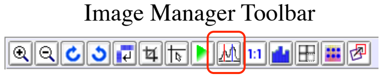
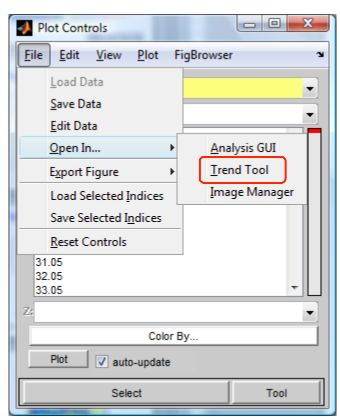


Image Manager Toolbar



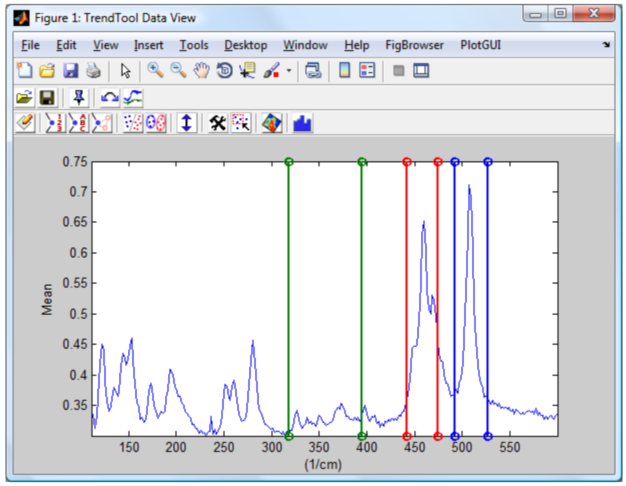
Plot Controls Window



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

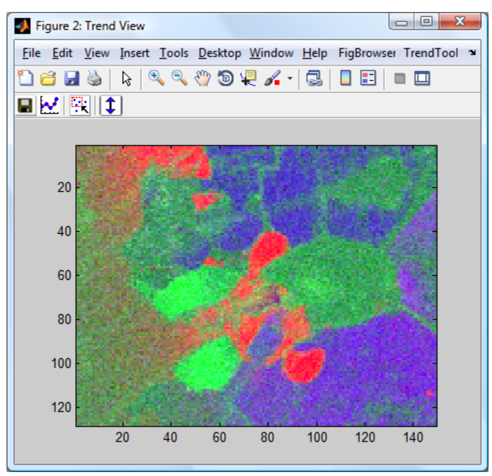


EIGENVECTOR RESEARCH INCORPORATED

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)



EIGENVECTOR RESEARCH INCORPORATED

TrendTool Analysis Modes

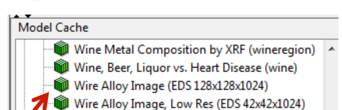
- **Height** – gives response at position (single marker)
- **Area** – gives integrated response between markers
- **Position** – gives position of peak response between markers
- **Width** – gives full width at half height between markers

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

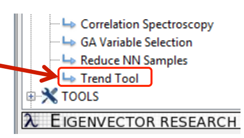
"Normalize to Region" to normalize all regions to the response of the selected region.

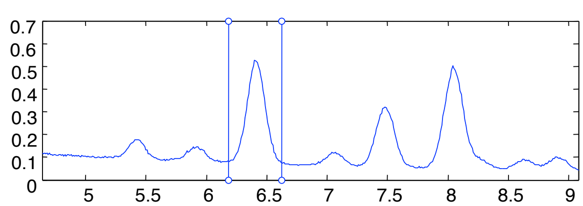
TrendTool Example

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)





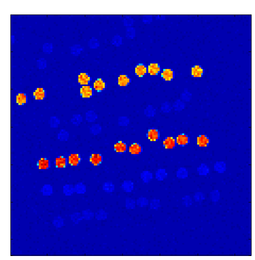
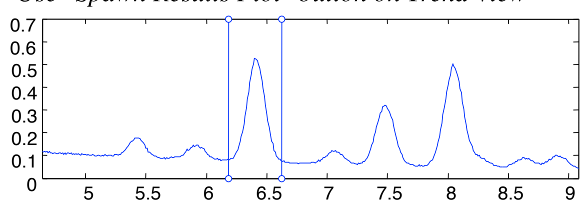
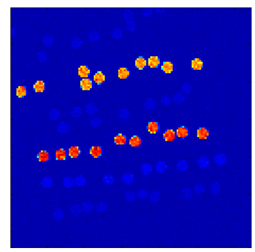


Image Exploration

- 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

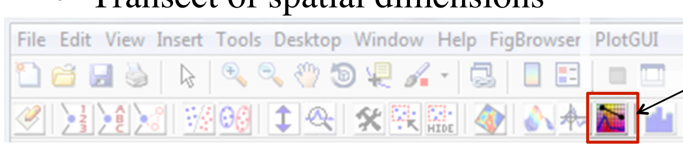




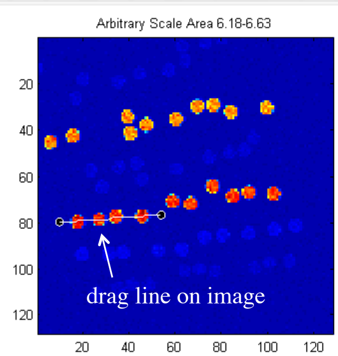
EIGENVECTOR RESEARCH INCORPORATED

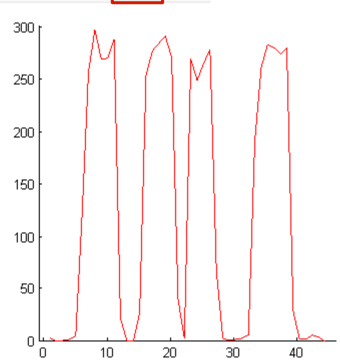
Cross-Section Tool

- Transect of spatial dimensions



Cross-section Tool



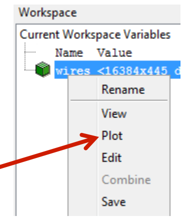


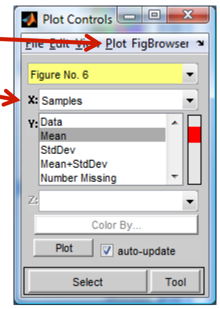
EIGENVECTOR RESEARCH INCORPORATED

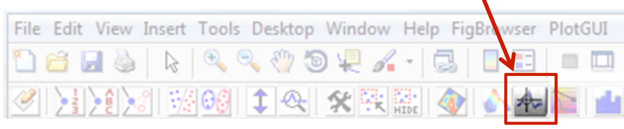
Drill Tool

- 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"



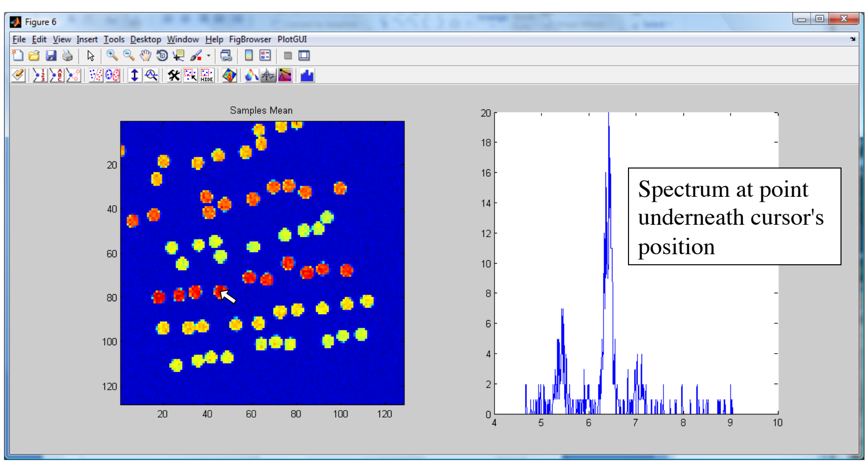




EIGENVECTOR RESEARCH INCORPORATED

23

Drill Tool



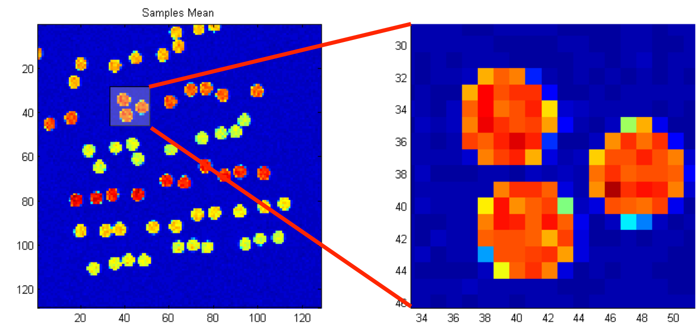
Double-click to view multiple spectra

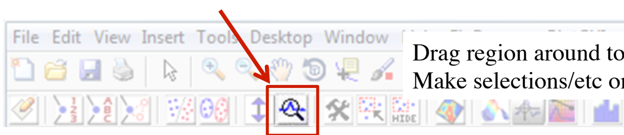
EIGENVECTOR RESEARCH INCORPORATED

24

Magnification Tool

- Show magnified view of image

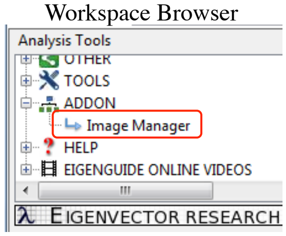




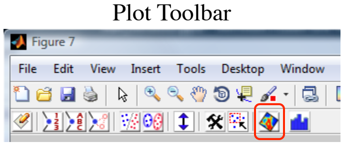
Drag region around to move magnifying glass
Make selections/etc on either image

Opening Image Manager

Workspace Browser



Plot Toolbar



Plot Controls Window

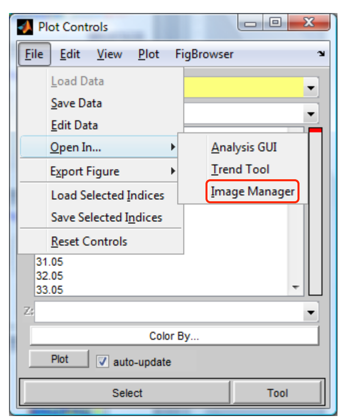
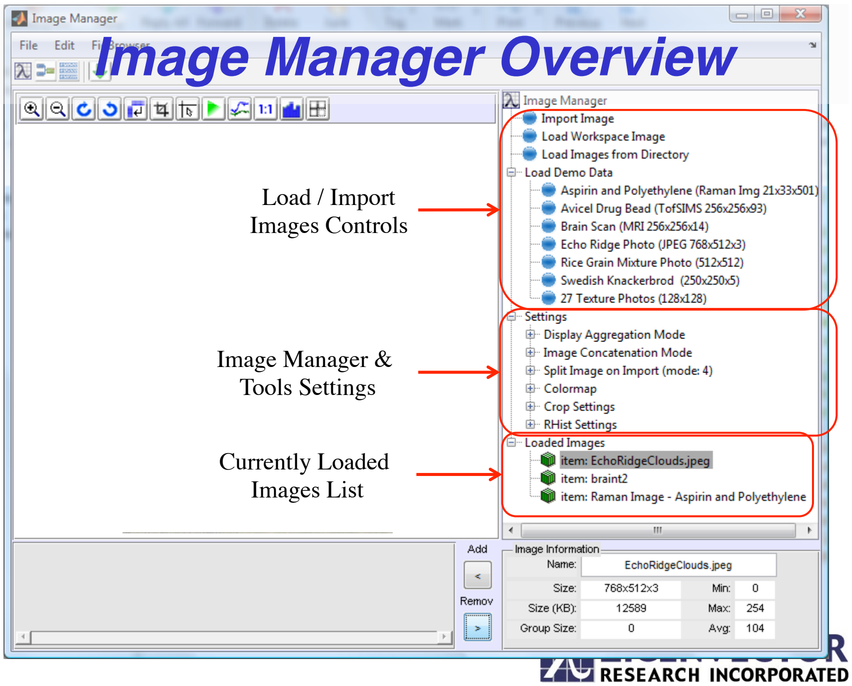
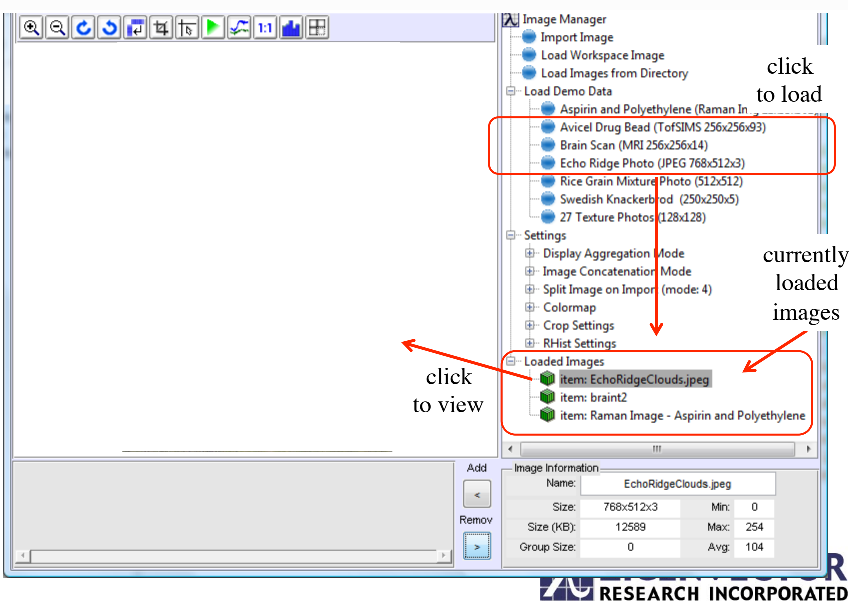


Image Manager Overview



27

Image Manager Overview

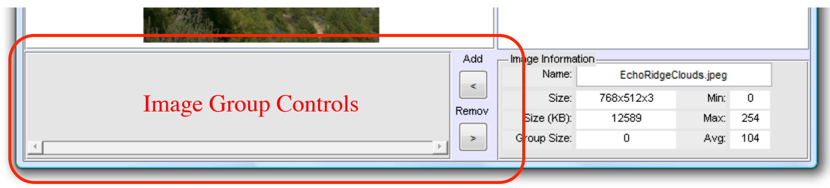


28

Image Groups

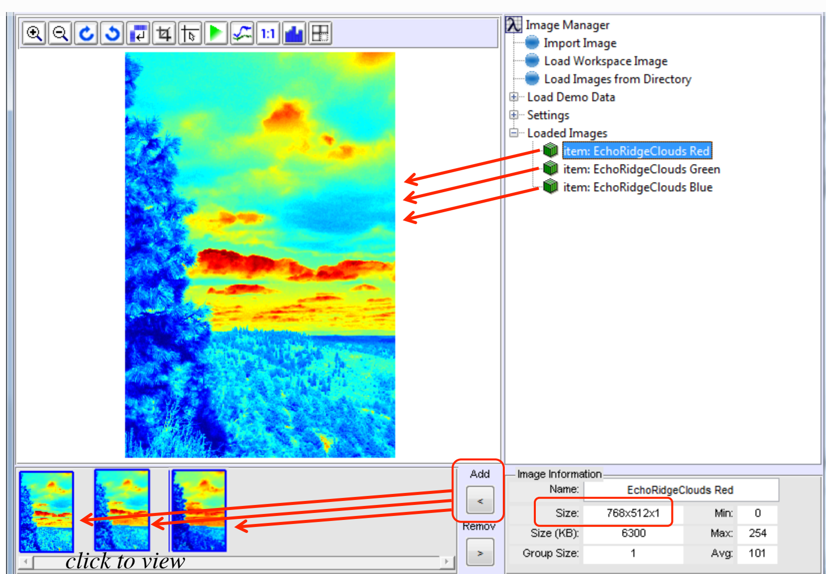
Grouping allows you to:

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

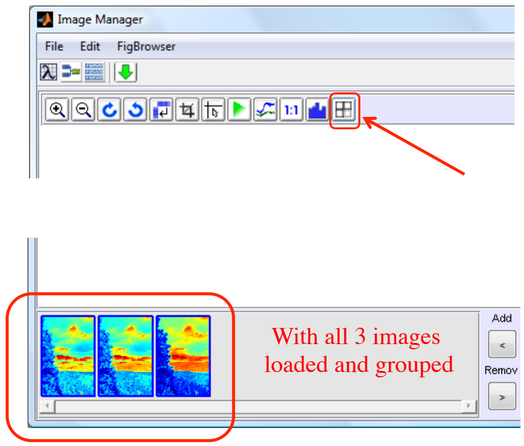


Example: combining three slabs of RGB image

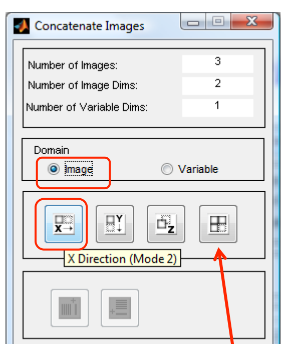
Image Groups

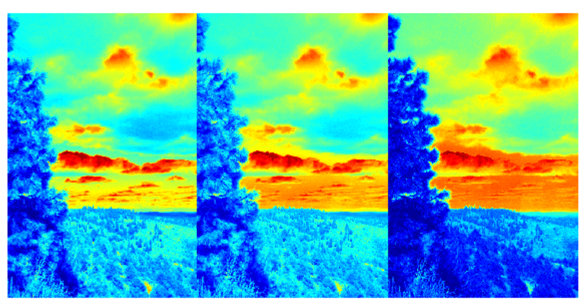


Concatenating Images



Concatenating Images: Spatial Domain

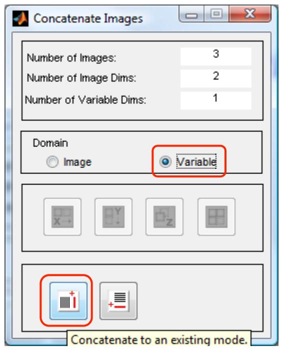




(768 x **1536**) x 1

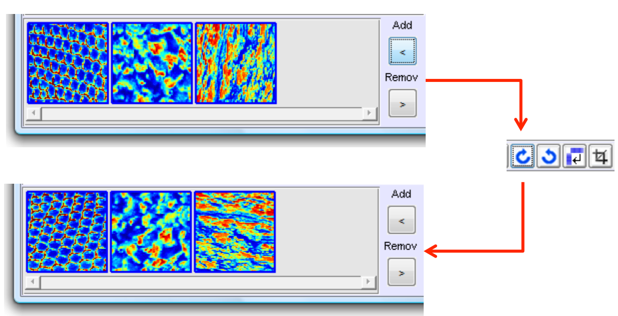
X, Y, Z, or tile…

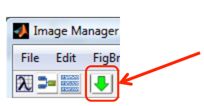
Concatenating Images: Variable Domain



(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.

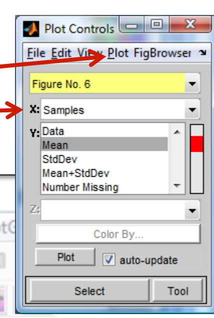
Particle Analysis

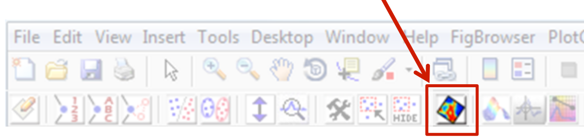
- Identify isolated regions (particles) in an image and give statistics on individual particles.
- Screen out particles and/or background.
- Create models based on particle statistics.
  - Particle outlier models (e.g. identify unusual particles)
  - Inferential models (e.g. drug activity based on particle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

- Based on long-established ImageJ platform.

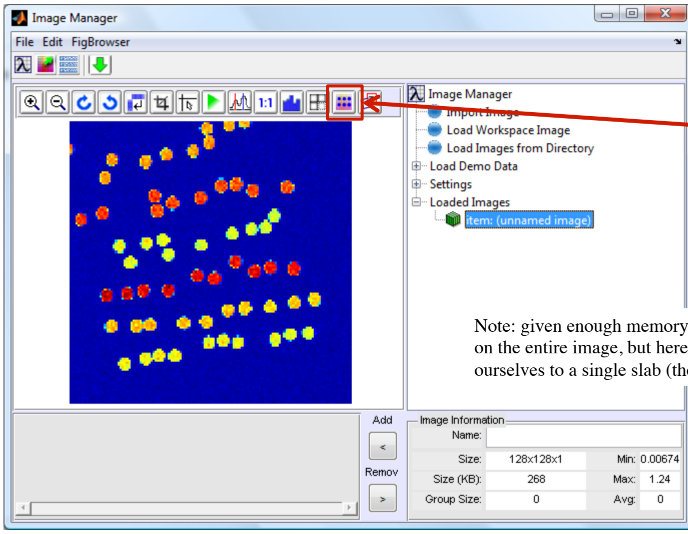
Particle Analysis Example

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"





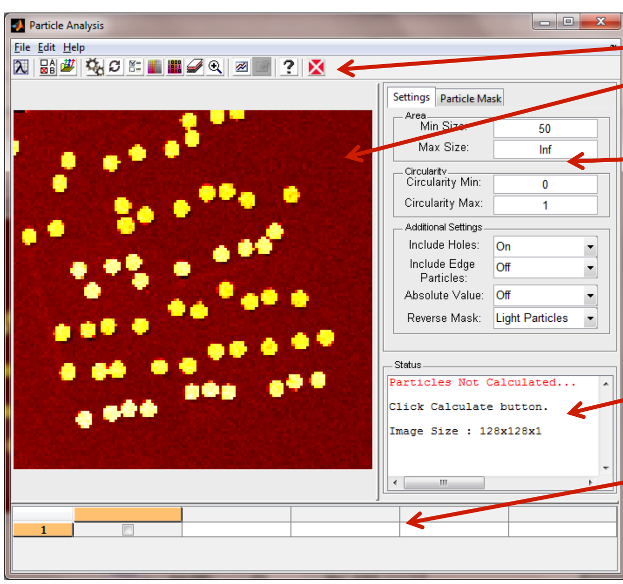
Particle Analysis Example



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).

EIGENVECTOR RESEARCH INCORPORATED



Actions Toolbar

Image with mask

Particle Analysis Settings

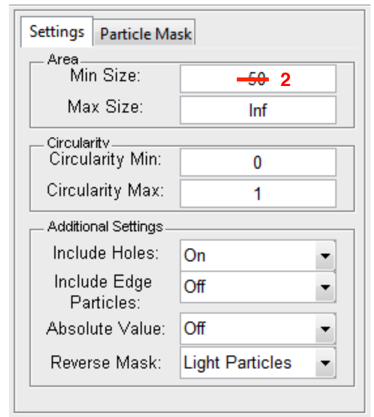
Summary of results

Table of located particles

EIGENVECTOR RESEARCH INCORPORATED

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).



EIGENVECTOR RESEARCH INCORPORATED

Particle Mask Settings

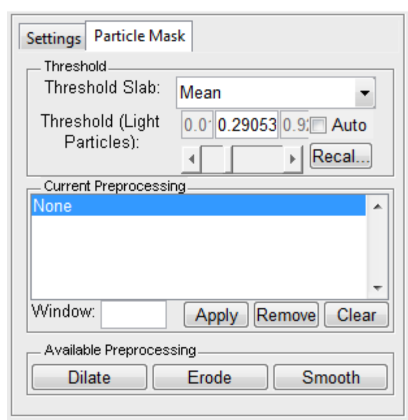
Adjusts which pixels are considered particles

•**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.

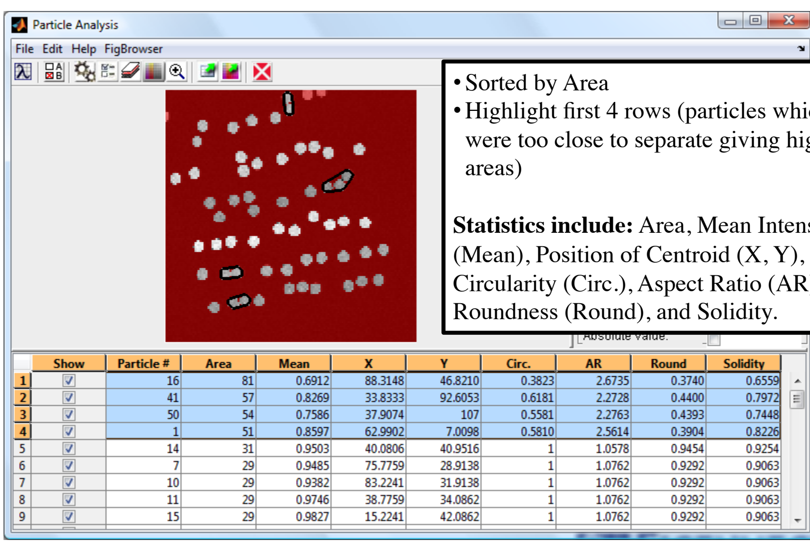


EIGENVECTOR RESEARCH INCORPORATED

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

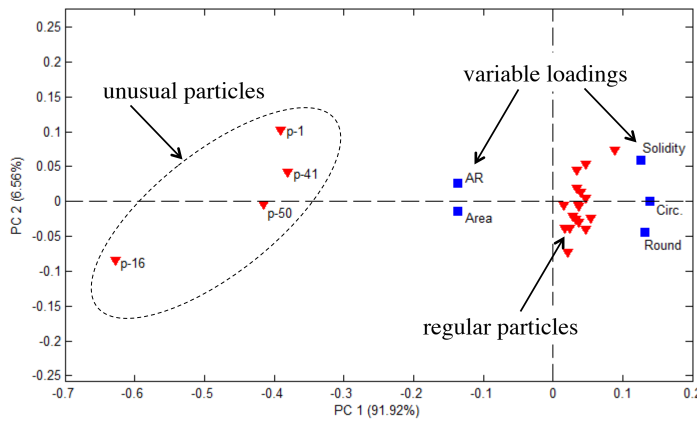


	Show	Particle #	Area	Mean	X	Y	Circ.	AR	Round	Solidity
1	☑	16	81	0.6912	88.3148	46.8210	0.3823	2.6735	0.3740	0.6559
2	☑	41	57	0.8269	33.8333	92.6053	0.6181	2.2728	0.4400	0.7972
3	☑	50	54	0.7586	37.9074	107	0.5581	2.2763	0.4393	0.7448
4	☑	1	51	0.8597	62.9902	7.0098	0.5810	2.5614	0.3904	0.8226
5	☑	14	31	0.9503	40.0806	40.9516	1	1.0578	0.9454	0.9254
6	☑	7	29	0.9485	75.7759	28.9138	1	1.0762	0.9292	0.9063
7	☑	10	29	0.9382	83.2241	31.9138	1	1.0762	0.9292	0.9063
8	☑	11	29	0.9746	38.7759	34.0862	1	1.0762	0.9292	0.9063
9	☑	15	29	0.9827	15.2241	42.0862	1	1.0762	0.9292	0.9063

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

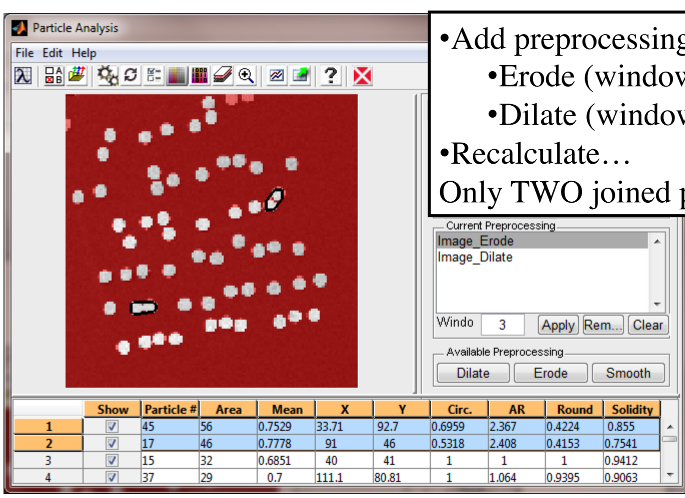
Statistics include: Area, Mean Intensity (Mean), Position of Centroid (X, Y), Circularity (Circ.), Aspect Ratio (AR), Roundness (Round), and Solidity.

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



	Show	Particle #	Area	Mean	X	Y	Circ.	AR	Round	Solidity
1	☑	45	56	0.7529	33.71	92.7	0.6959	2.367	0.4224	0.855
2	☑	17	46	0.7778	91	46	0.5318	2.408	0.4153	0.7541
3	☑	15	32	0.6851	40	41	1	1	1	0.9412
4	☑	37	29	0.7	111.1	80.81	1	1.064	0.9395	0.9063

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

Only TWO joined particles

Image-Oriented Preprocessing

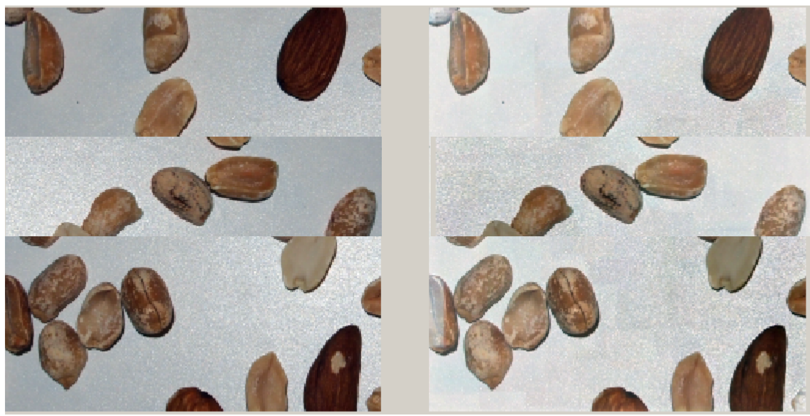
- Image-specific preprocessing operates in pixel-space and are either Intensity or Binary based
- Intensity-Based Image Correction:
  - *Background Subtraction (Flatfield)*: Rolling-ball background subtraction for images.
  - *Min*: Min value over neighboring pixels. (filter out high-value pixels)
  - *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

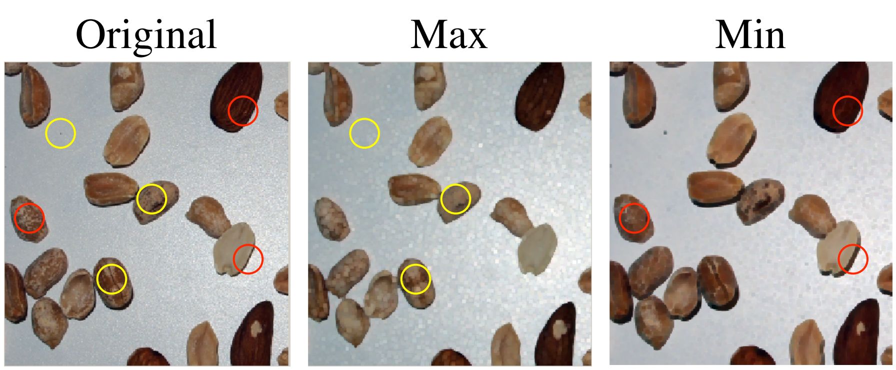
- Binary-Based Image Correction
  - *Dilate*: Perform dilation on a binary image.
  - *Erode*: Perform erosion on a binary image.
  - *Close (Dilate+Erode)*: Perform dilation followed by erosion on a binary image.
  - *Open (Erode+Dilate)*: Perform erosion followed by dilation on a binary image.

- 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

Background Subtraction (Flat-field)



Max & Min Preprocessing



Mean & Median Preprocessing

Original | Mean | Median

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

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

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

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

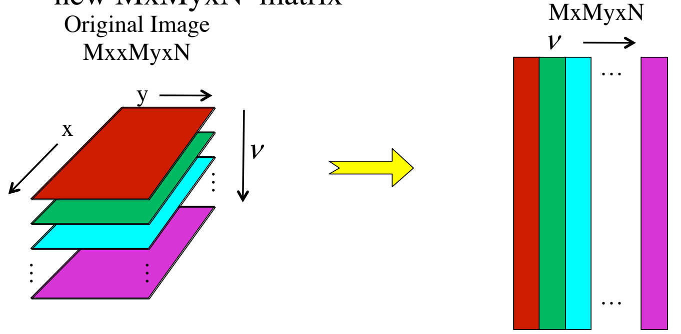
  Where $R \leq \min\{M,N\}$, and the $\mathbf{t}_k\mathbf{p}_k^T$ pairs are ordered by the amount of variance captured.
- Generally, the model is truncated to *K* PCs, leaving some small amount of variance in a residual matrix **E**:

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

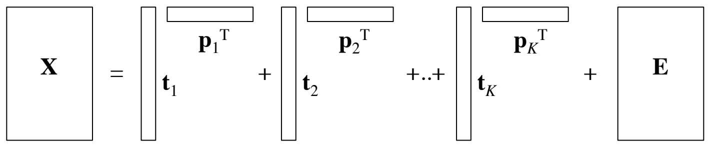
- where **T** is *M*x*K* and **P** is *N*x*K*.

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

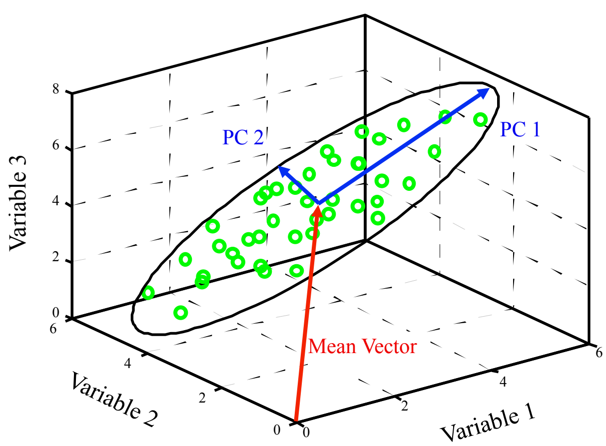


Properties of PCA



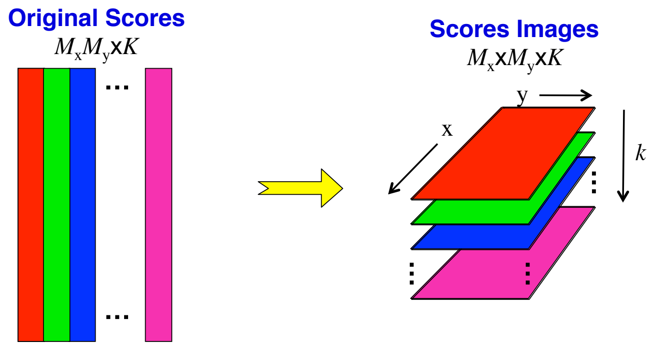
- $\mathbf{t}_k$,$\mathbf{p}_k$ ordered by amount of *variance captured*
  - λ_k are the eigenvalues of $\mathbf{X}^T\mathbf{X} \rightarrow \mathbf{X}^T\mathbf{X}\mathbf{p}_k = \lambda_k\mathbf{p}_k$
  - λ_k are $\propto$ variance captured
- $\mathbf{t}_k$ (*scores*) form an orthogonal set $\mathbf{T}_K$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

PCA Graphically



Reshape Scores To Images

- PCA gives scores **T** (MxK) which is reshaped to scores images (M_xxM_yxK)
  - each score vector is a M_xxM_y scores image



EIGENVECTOR RESEARCH INCORPORATED

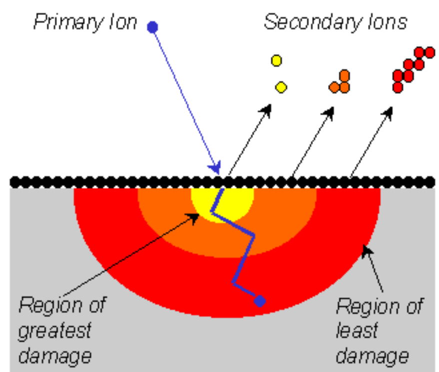
Plots / Images for PCA

- scores and loadings plots are interpreted in pairs
  - plot $\mathbf{t}_k$ vs sample number
    - find relationship between *samples* → *pixels*
    - each M_xM_yx1 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

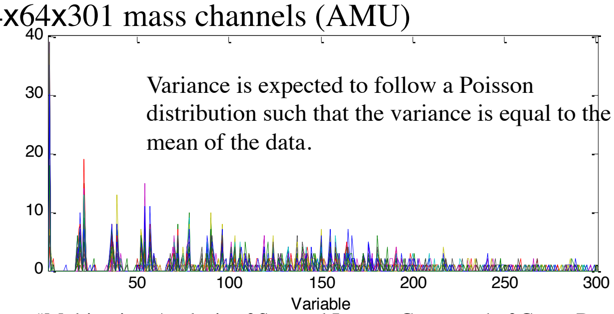
TOF-SIMS of PMMA and Deuterated Polystyrene

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

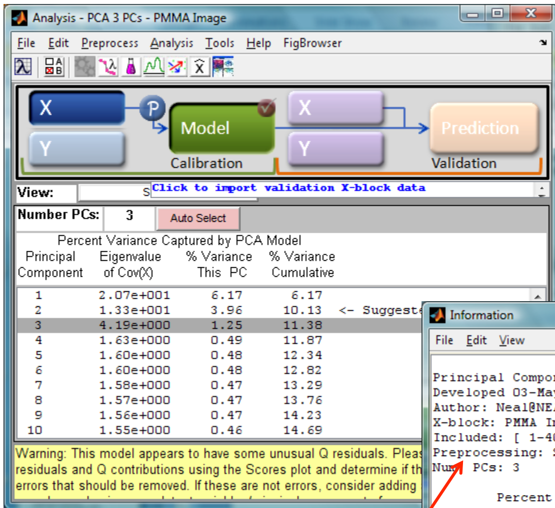


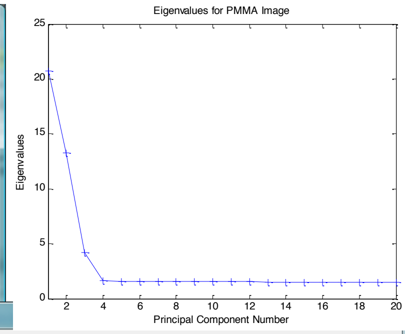
Example Data

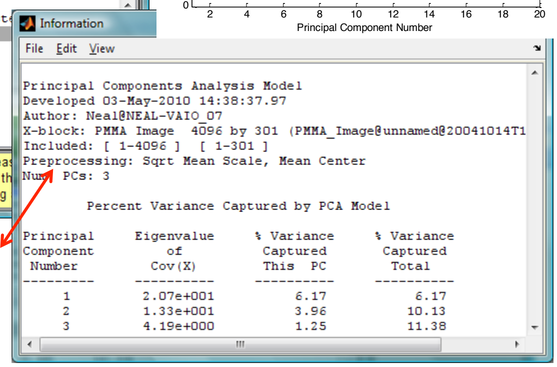
- Data is positive SIMS spectrum at each pixel (point) on a 64x64 grid
- 64x64x301 mass channels (AMU)



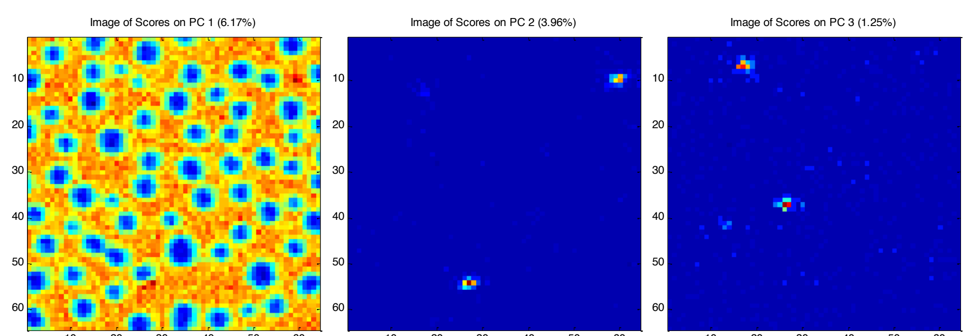
M.R. Keenan, "Multivariate Analysis of Spectral Images Composed of Count Data," in *Techniques and Applications of Hyperspectral Image Analysis*, H. F. Grahn and P. Geladi, eds. (John Wiley & Sons, West Sussex, England), 89-126, 2007.







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 **E**, $\mathbf{e}_m$, and a row of **X**, $\mathbf{x}_m$, $m = 1, \ldots, M$

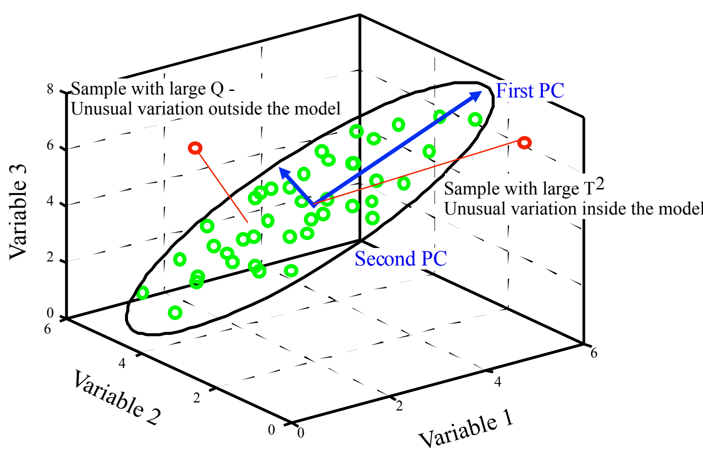
$$Q_m = \mathbf{e}_m \mathbf{e}_m^T = \mathbf{x}_m (\mathbf{I}-\mathbf{P}_K \mathbf{P}_K^T)\mathbf{x}_m^T$$

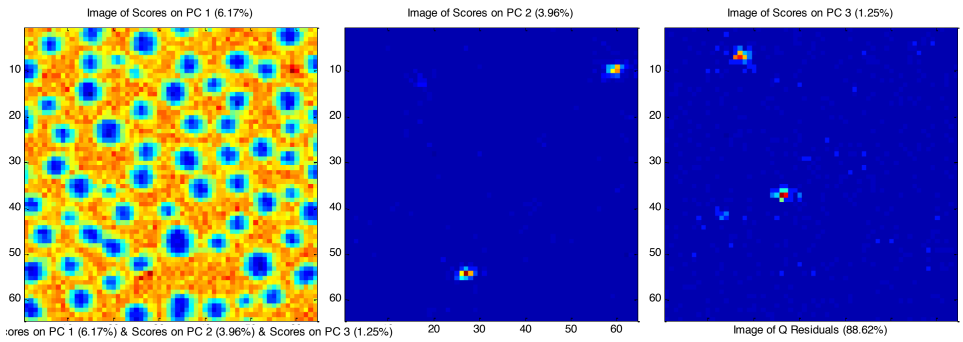
  - Hotelling's T^2 statistic
    - for a row of $\mathbf{T}_K$, $\mathbf{t}_m$, and KxK diagonal matrix λ

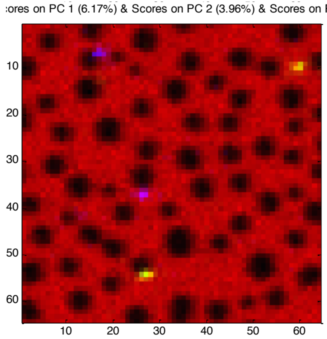
$$T_m^2 = \mathbf{t}_m \lambda^{-1} \mathbf{t}_m^T = \mathbf{x}_m \mathbf{P}_K \lambda^{-1} \mathbf{P}_K^T \mathbf{x}_m^T$$

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

Geometry of Q and T^2

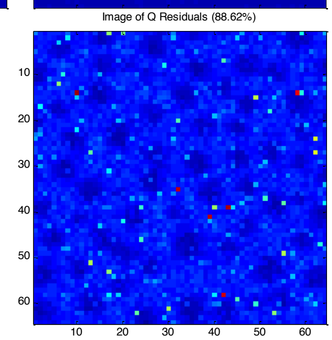




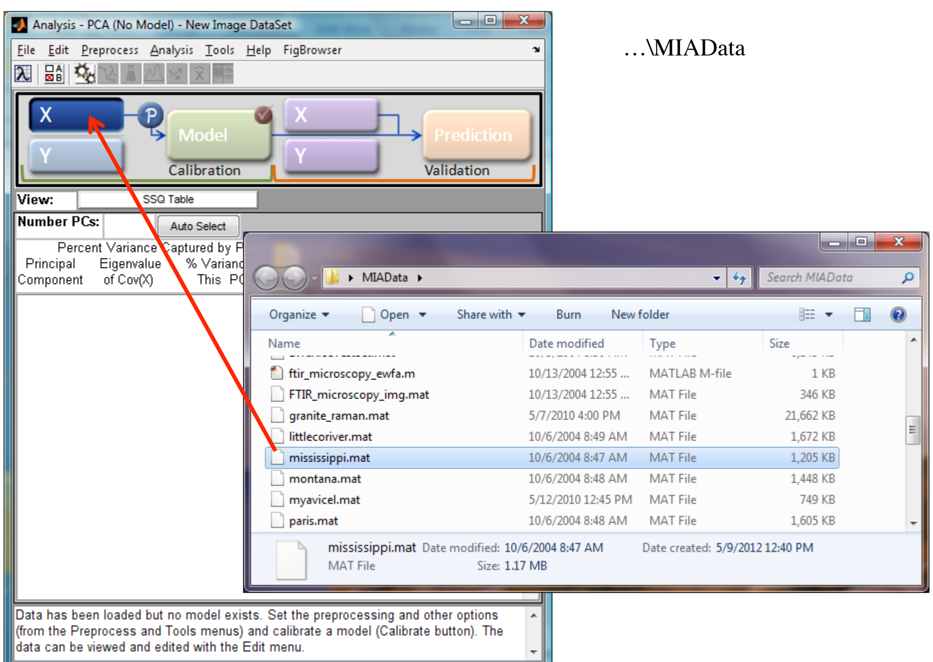


RGB of scores on PCs 1, 2 and 3

Q residuals

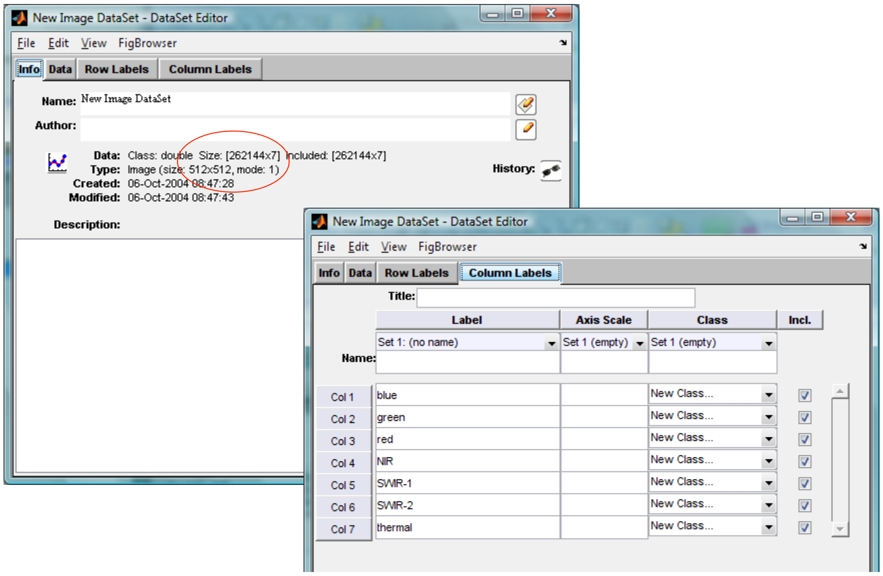


EIGENVECTOR RESEARCH INCORPORATED



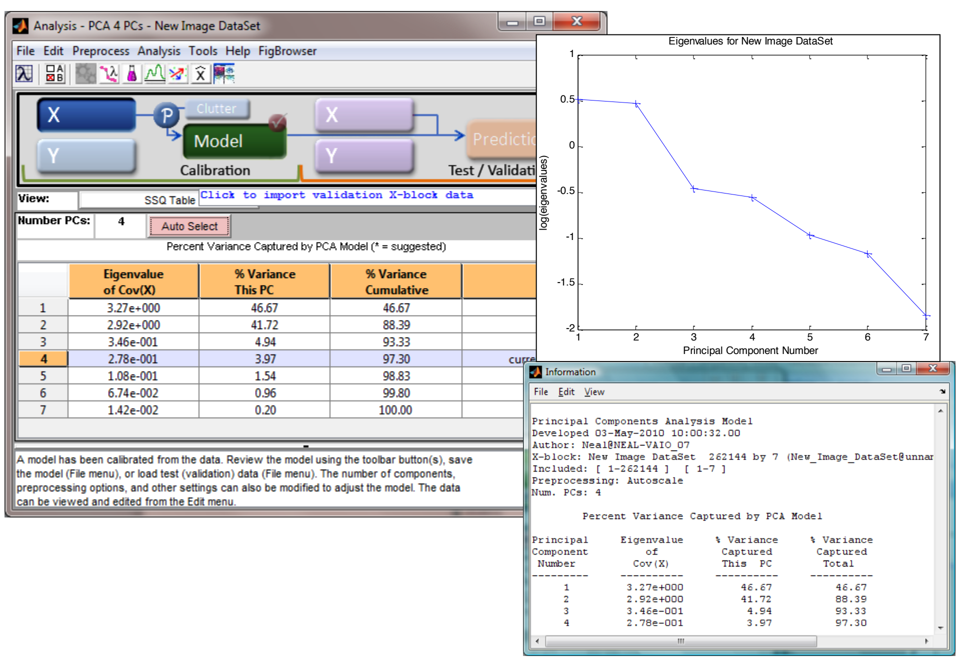
…\MIAData

EIGENVECTOR RESEARCH INCORPORATED

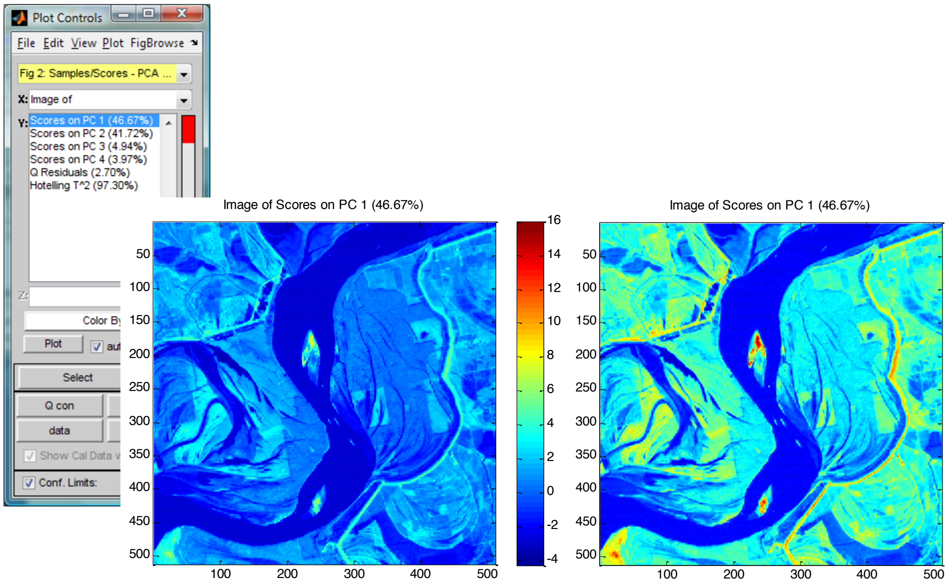


EIGENVECTOR RESEARCH INCORPORATED

67

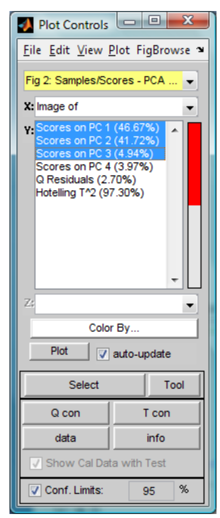


EIGENVECTOR RESEARCH INCORPORATED

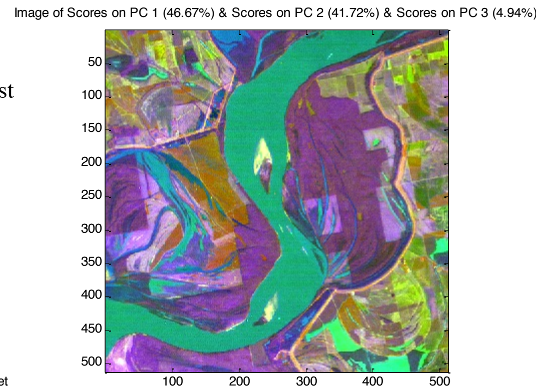


Creating Color Images

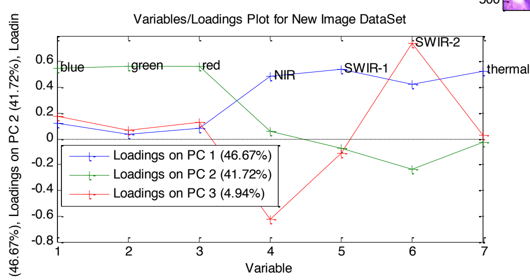
- Images are made of three colors: red, green and blue
  - e.g., scaled to integers 0-255 for 8-bit color
- Scores can be used to define the colors
  - PC 1 = red, PC 2 = green, PC 3 = blue

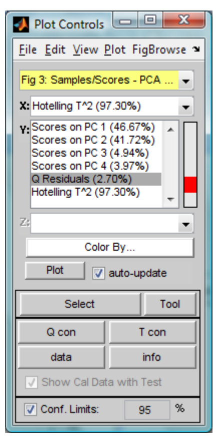


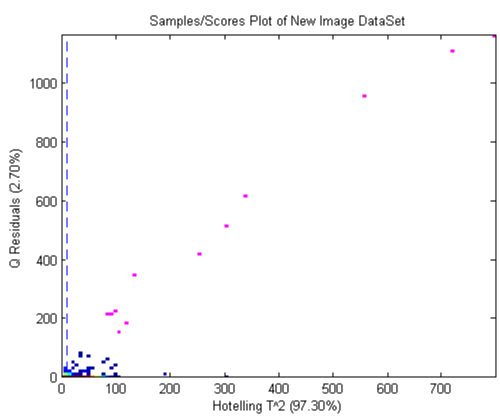
Scores
View:Auto Contrast



Loadings
View:Labels:Set 1





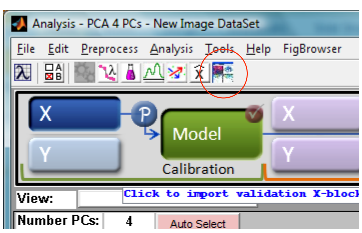


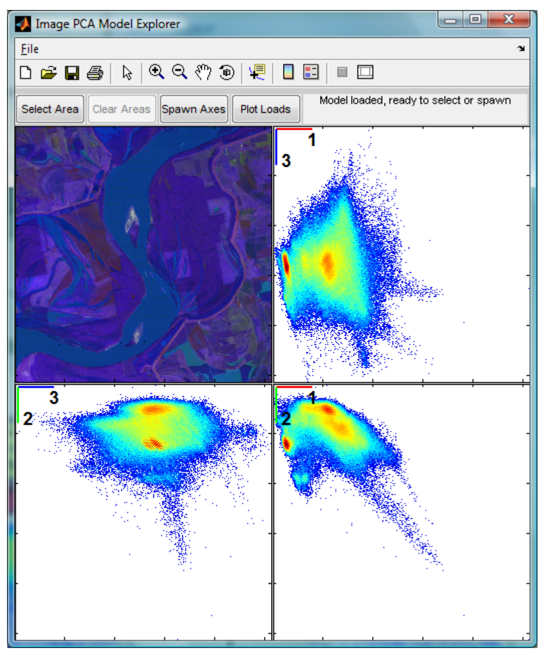
pixels with high Q and T2 have been selected

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

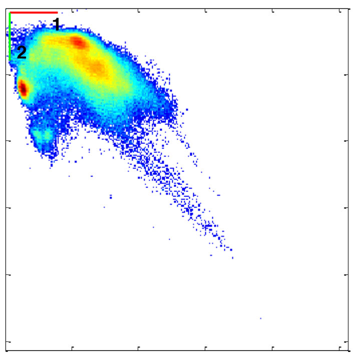
EIGENVECTOR RESEARCH INCORPORATED

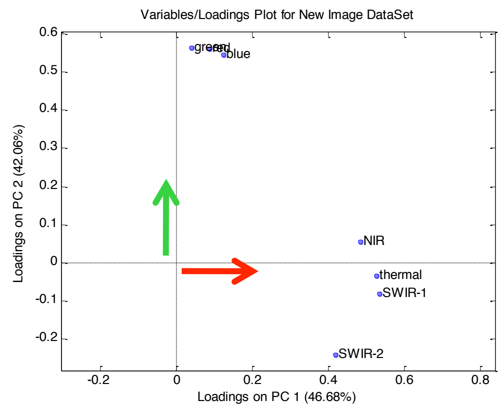


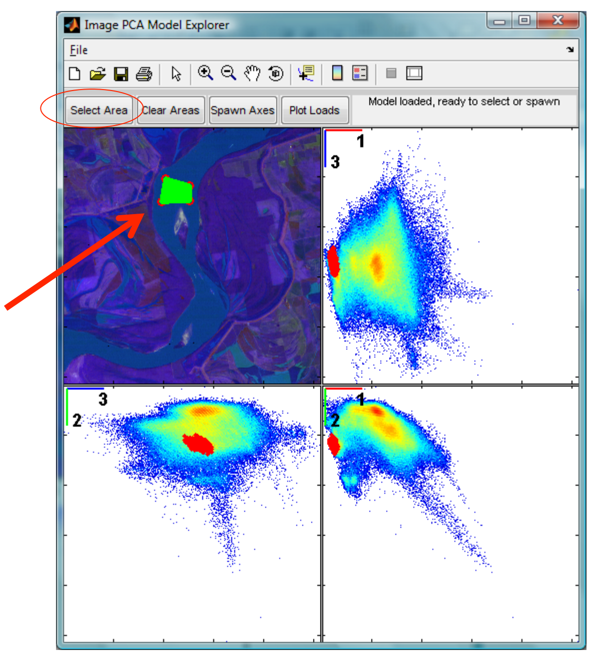


EIGENVECTOR RESEARCH INCORPORATED

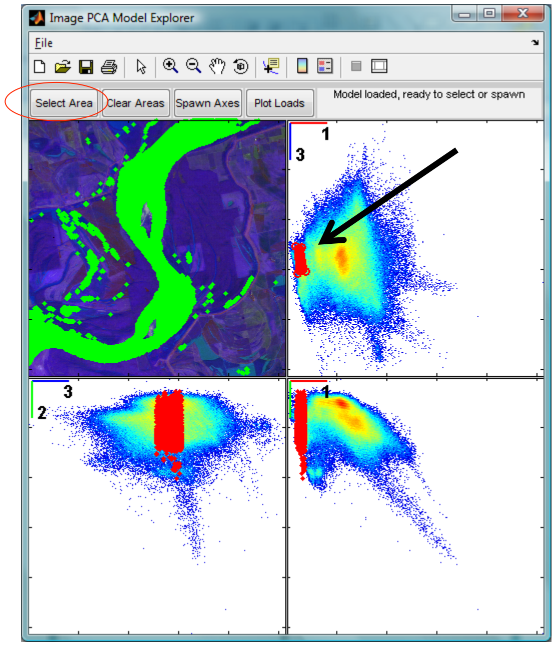
Scores and Loadings







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



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

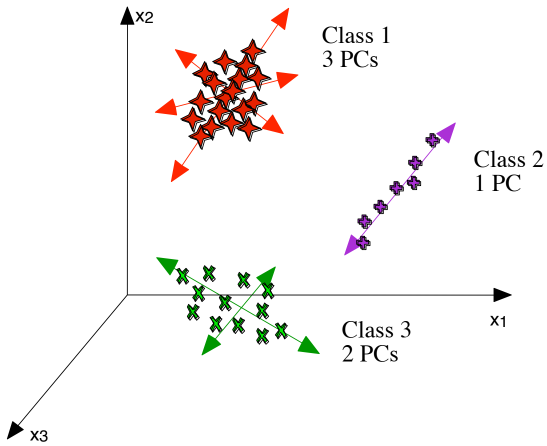
EIGENVECTOR RESEARCH INCORPORATED

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

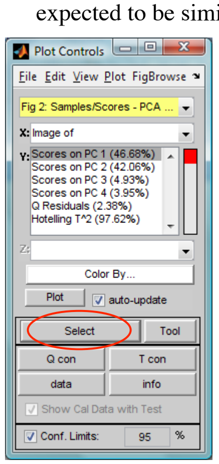
EIGENVECTOR RESEARCH INCORPORATED

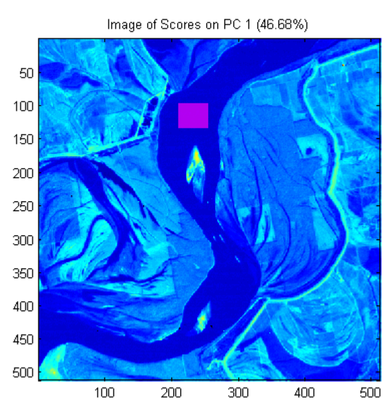
A SIMCA Model



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.

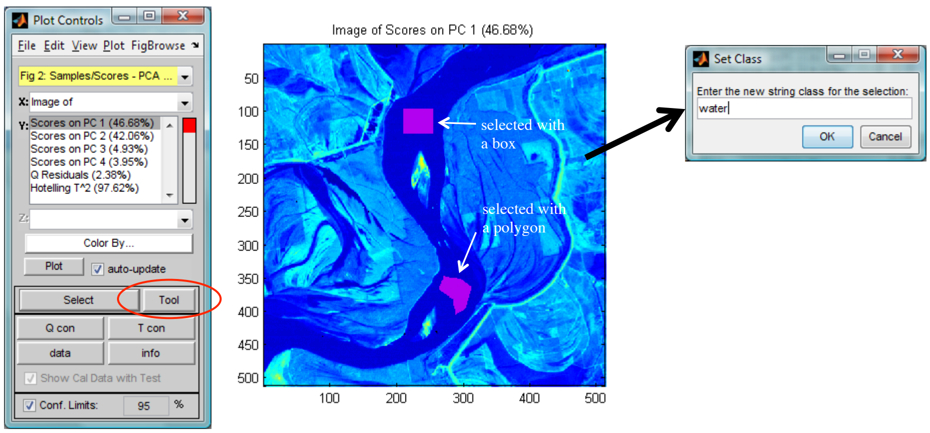




SIMCA Example

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

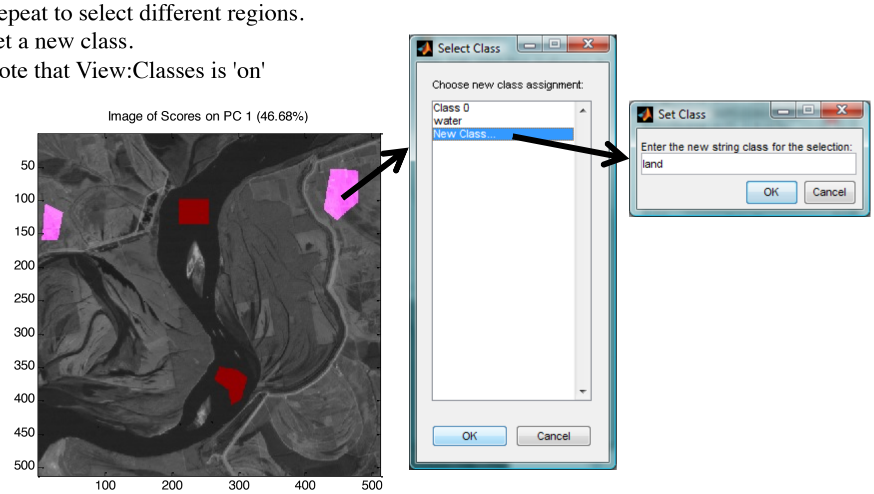
Edit > Set Class of Selection



EIGENVECTOR RESEARCH INCORPORATED

SIMCA Example

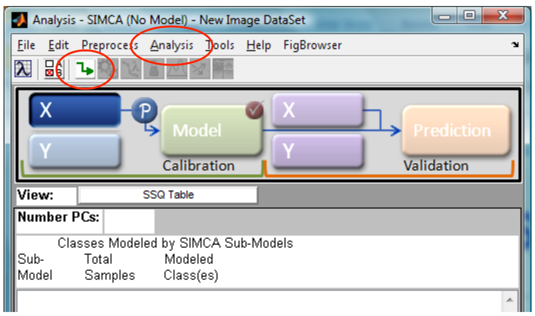
- Repeat to select different regions.
- Set a new class.
- Note that View:Classes is 'on'

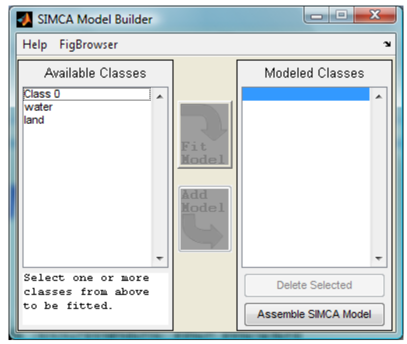


EIGENVECTOR RESEARCH INCORPORATED

SIMCA Model Builder

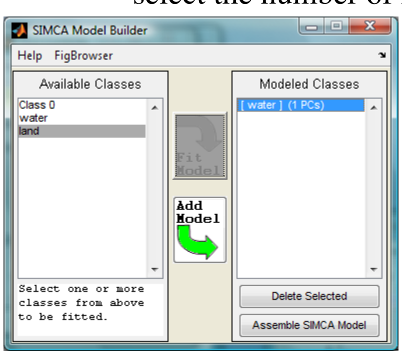
- SIMCA requires a selection of classes to be modeled and then assembles the model
  - Analysis:SIMCA

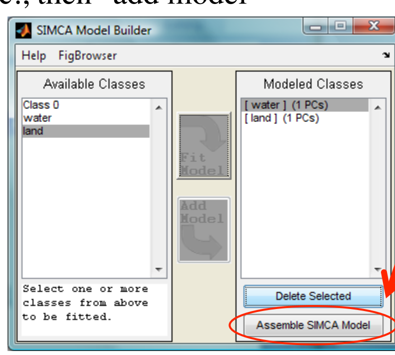




Model of Each Class

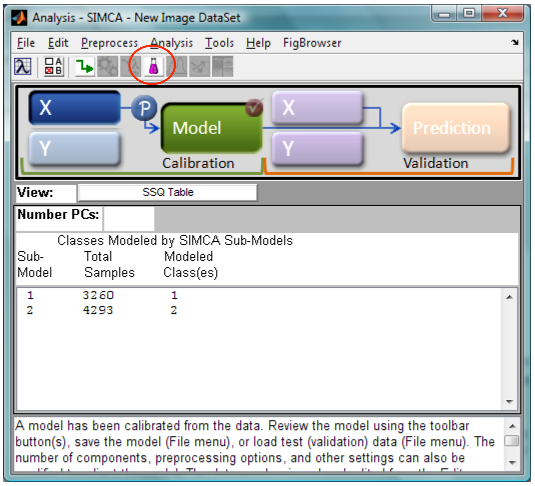
- Each class is modeled using PCA
  - highlight a class and then "fit model"
  - select the number of PCs, etc., then "add model"





SIMCA Example

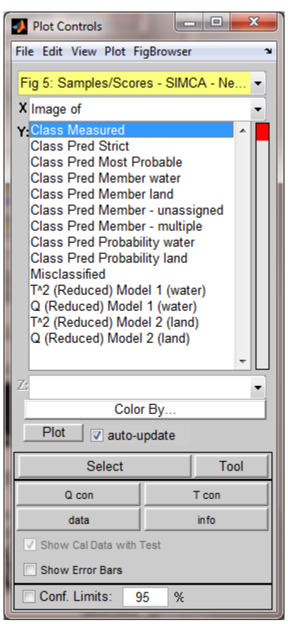
- The SIMCA model consists of two PCA models



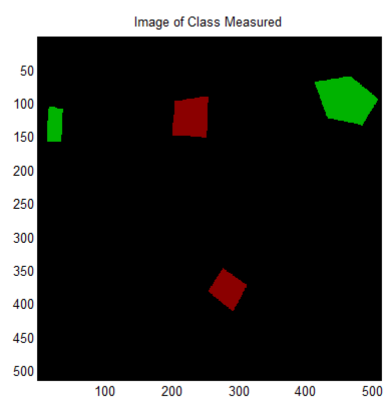
- Data from the entire image will be projected onto each PCA model.
- Scores, Q and T^2 are calculated for each model and it is determined which model the data is closest to.
- Click the scores button to examine the images.

EIGENVECTOR RESEARCH INCORPORATED

SIMCA Model Predictions



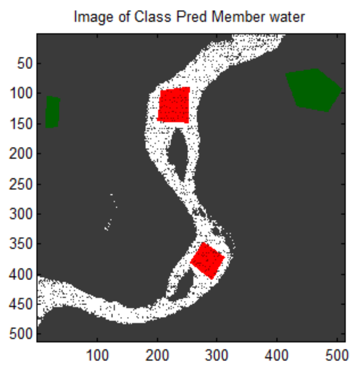
- "Class Measured" = where the classes were selected.
- "Reduced" means that the statistic was normalized by the limit of the corresponding statistic (e.g., to the 95% CL).

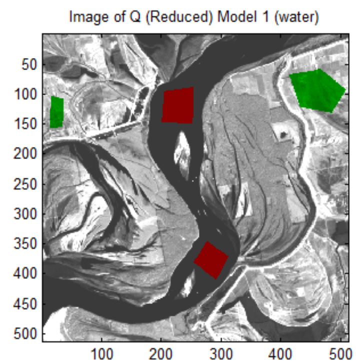


EIGENVECTOR RESEARCH INCORPORATED

Model 1 Predictions

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

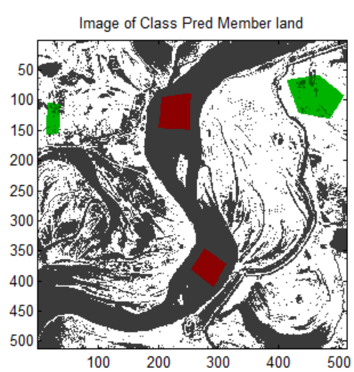


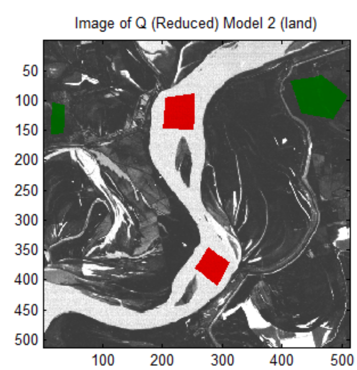


EIGENVECTOR RESEARCH INCORPORATED

Model 2 Predictions

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

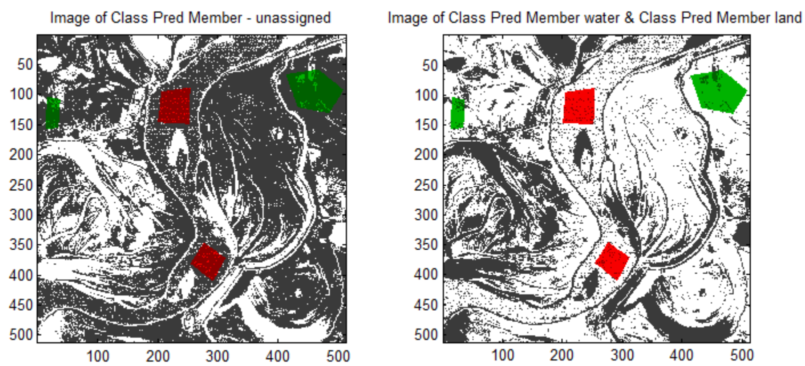




EIGENVECTOR RESEARCH INCORPORATED

In Model and Not-In-Any Model

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



"Strict" Class Predictions

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

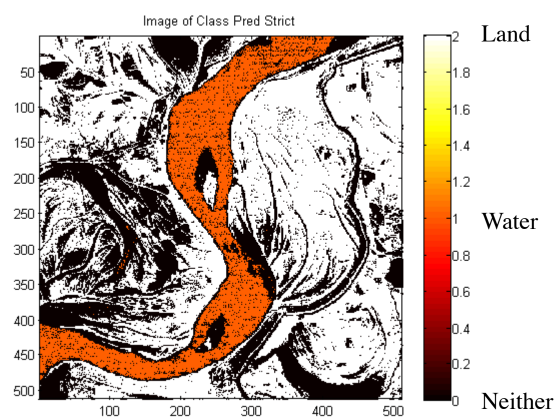
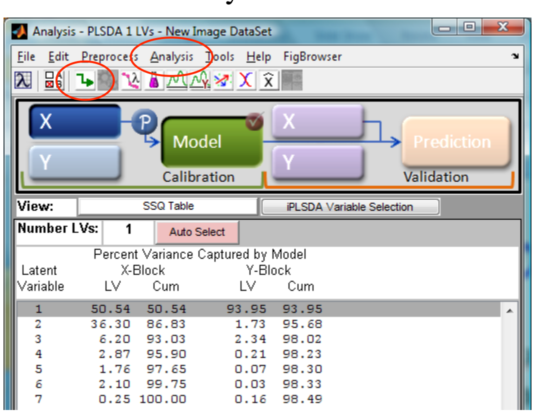


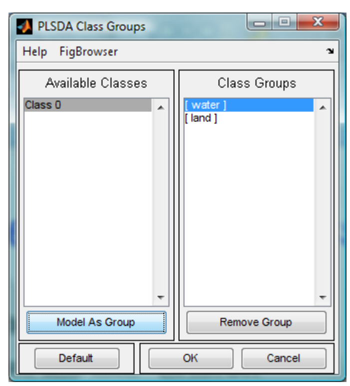
Image PCA Conclusions

- 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

PLSDA Model Builder

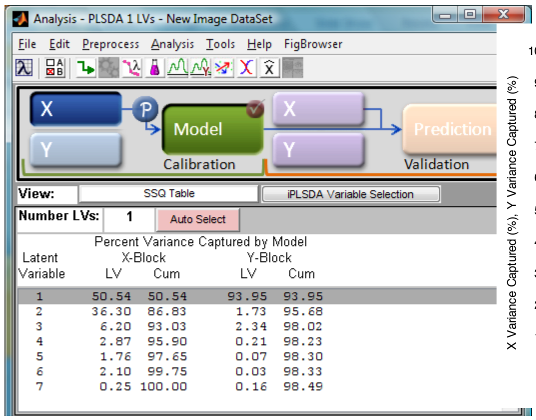
- PLS discriminant analysis requires a selection of classes to be modeled
  - Analysis:PLSDA

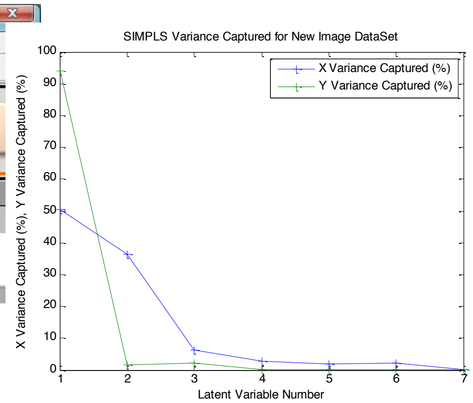




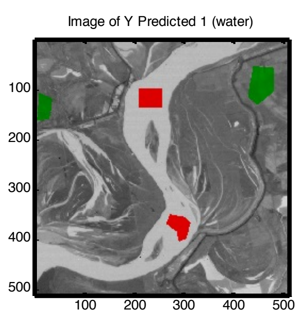
PLSDA Maximizes Class Separation on a PLS Model

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

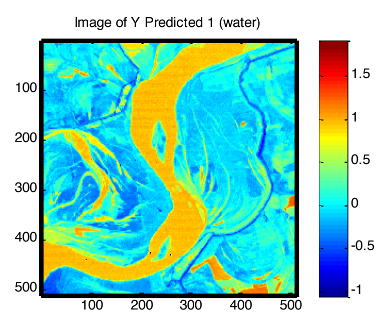


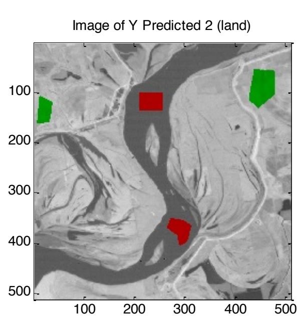


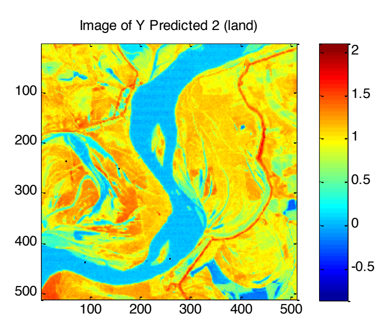
EIGENVECTOR RESEARCH INCORPORATED



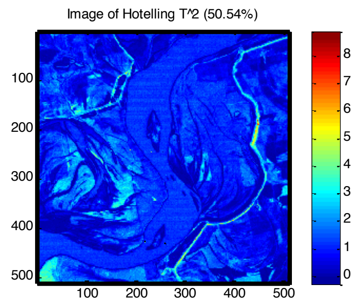
- Data from the entire image are projected onto the PLSDA model.
- Light shows high predictions on each class.
- Click the scores button to examine the images.
- View:Classes (uncheck Set 1)

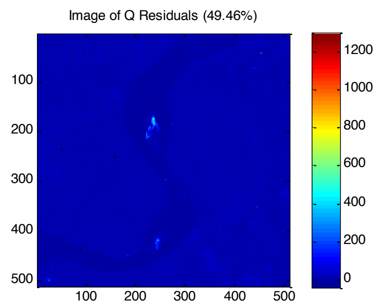




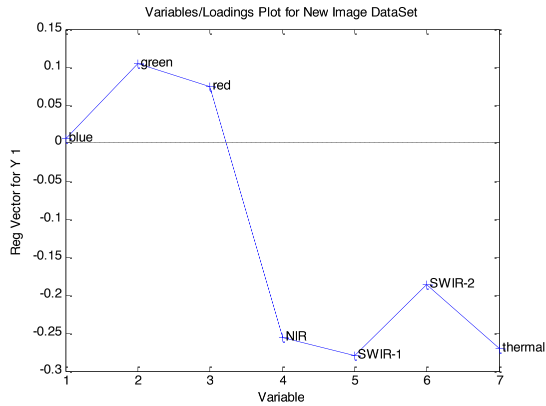


EIGENVECTOR RESEARCH INCORPORATED



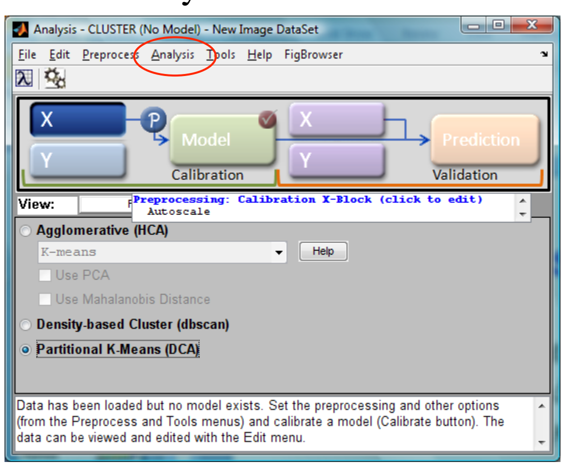


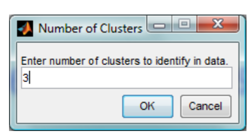
- Inspect T^2 and Q
- Regression vector suggests that green and red increase relative to IR channels for water relative to land



Cluster Analysis

- Analysis:Cluster





Results for 3 and 5 Clusters

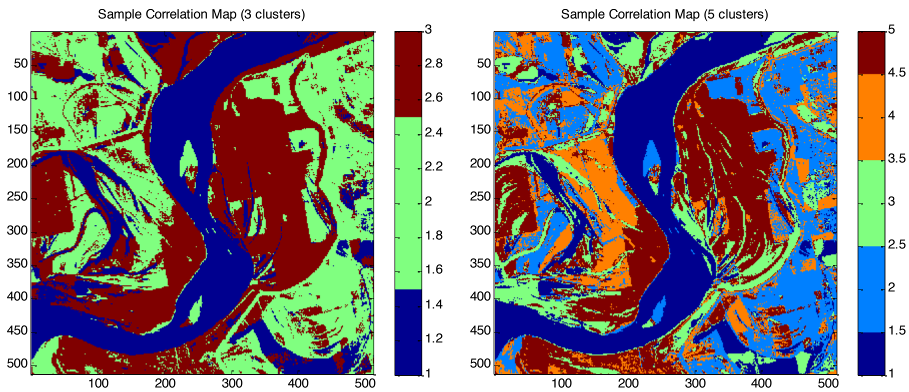


Image PLSDA and Clustering Conclusions

- 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

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

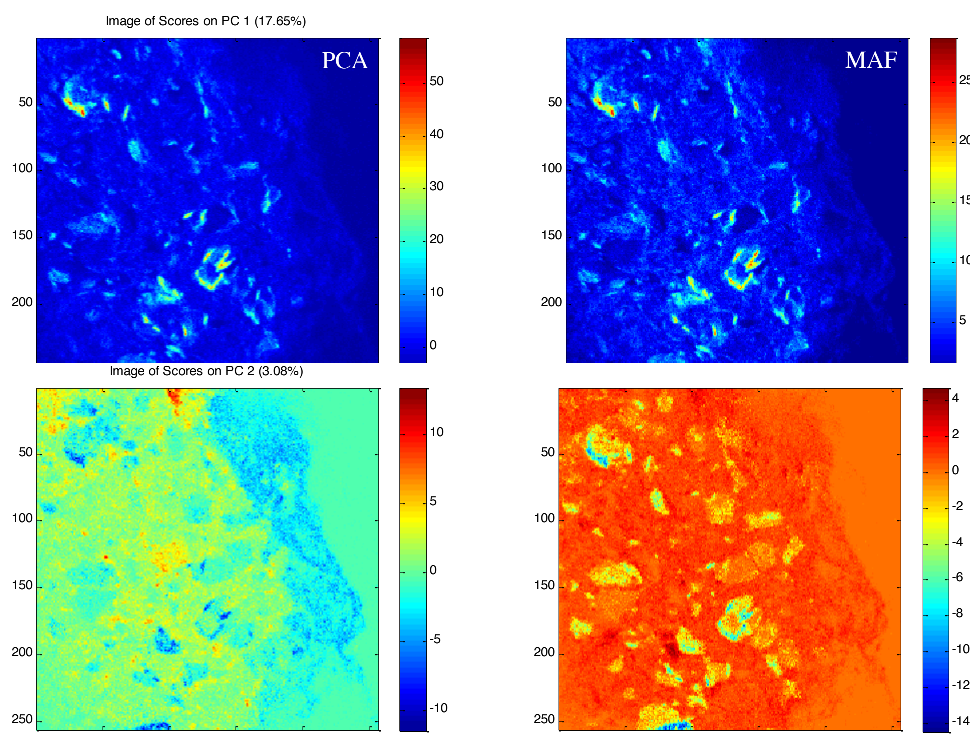
EIGENVECTOR RESEARCH INCORPORATED

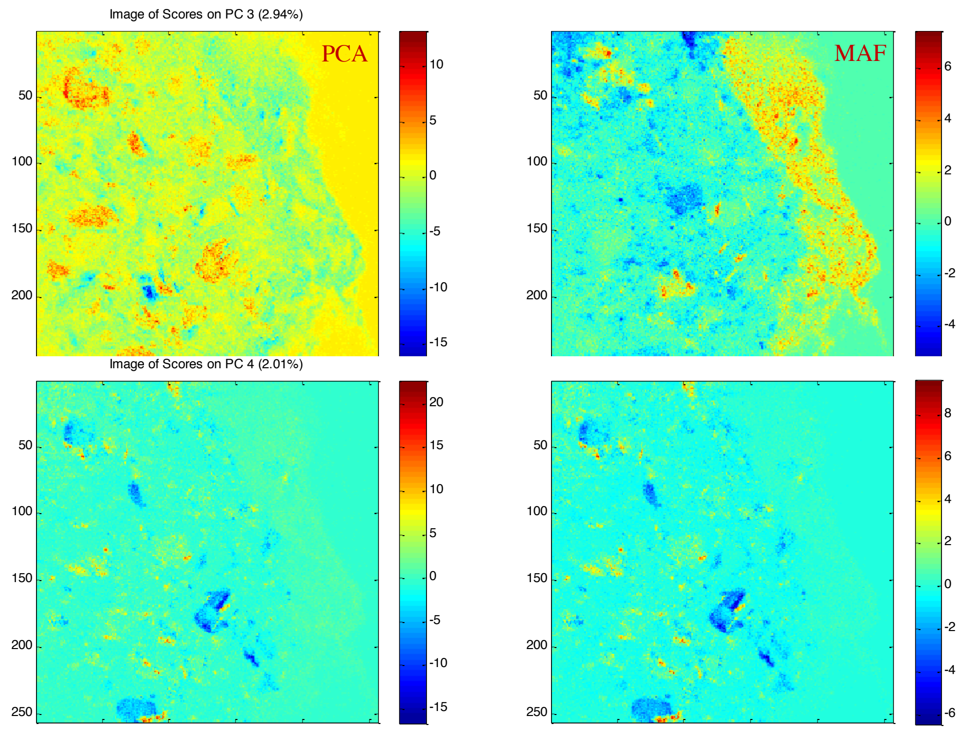
Maximum Autocorrelation Factors for Multivariate Images

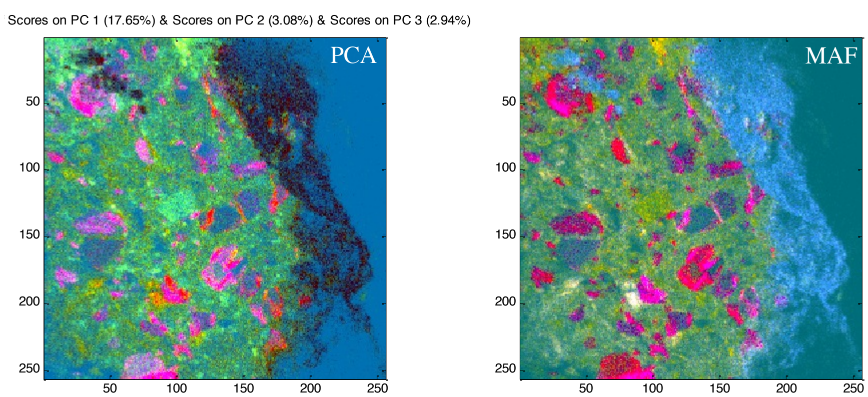
- For MNF, the clutter was intra-class variance
- For MAF, the clutter is the first spatial difference
  - the first difference should be high on edges and just noise w/in clusters
  - the result is the same generalized eigenvector problem as MNF with different clutter $\mathbf{\Sigma}_C$

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







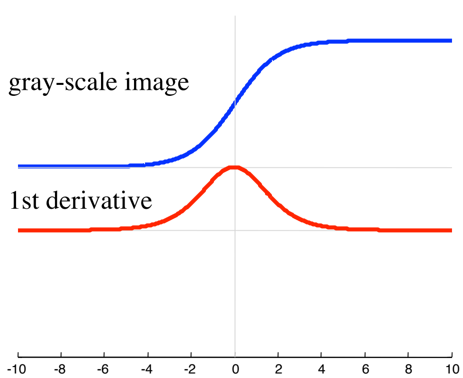
RGB images after auto-contrasting

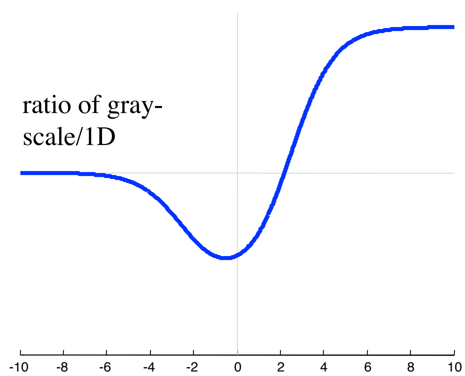
Maximum Difference Factors MDF

- For MNF and MAF $\boldsymbol{\Sigma}_X$ was the covariance of the image
- For MAF $\boldsymbol{\Sigma}_C$ was the covariance of the first spatial difference and in MNF it was estimated from intra-class variance
- For MDF $\boldsymbol{\Sigma}_X$ is the covariance of the first spatial derivative of image, and $\boldsymbol{\Sigma}_C$ is the covariance of the second spatial derivative
  - the result is multivariate edge detection
  - often show magnitudes sqrt(dx^2+dy^2)

EIGENVECTOR RESEARCH INCORPORATED

MAF

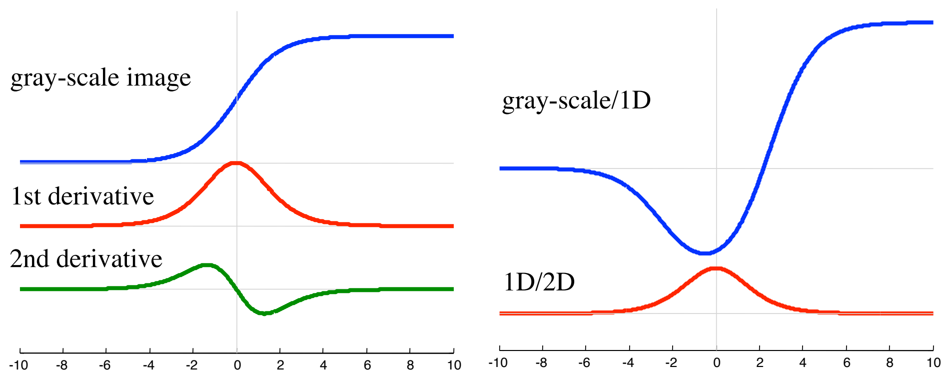




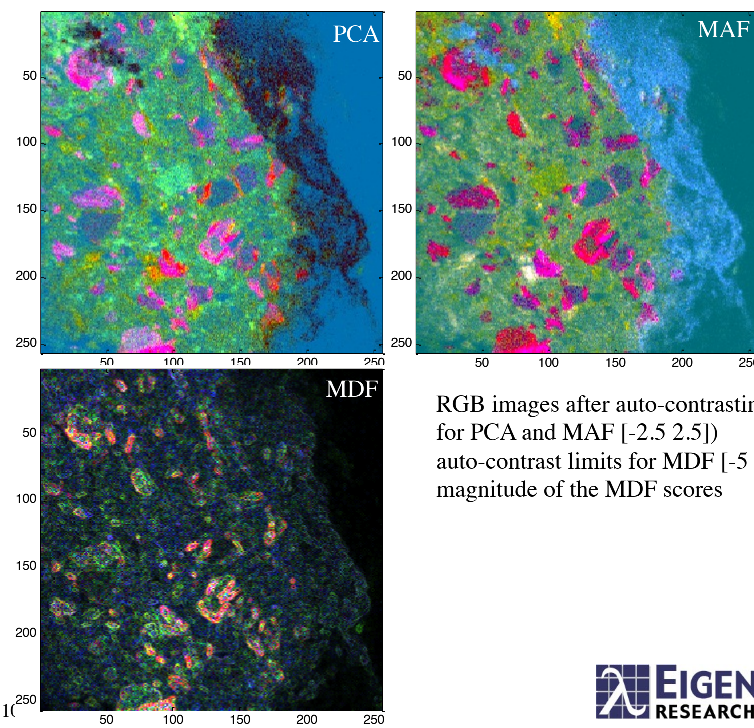
MAF finds locations in the image where the ratio of gray-scale to first derivative is a maximum

EIGENVECTOR RESEARCH INCORPORATED

MDF



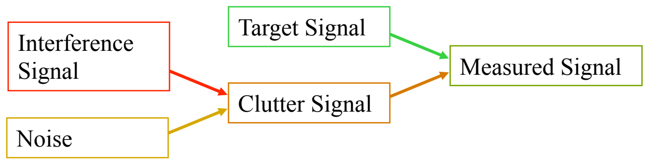
MDF finds locations in the image where the ratio of first to second derivative is a maximum



RGB images after auto-contrasting (default for PCA and MAF [-2.5 2.5])
auto-contrast limits for MDF [-5 5] for magnitude of the MDF scores

Measured Signal Includes Clutter

- Clutter is present in all measurements
  - X-block, Y-block



  - Use knowledge of physics and chemistry to create a linear relationship
    - non-linearity w/in X-block adds factors in X
    - non-linearity between X- and Y-blocks adds error

Why is Clutter Bad?

- Attempt to maximize S/C via pre-processing or the model e.g., MAF
- Methods that don't remove net analyte signal (NAS) are preferable
  - NAS is the portion of spectrum $\mathbf{s}_i$ unique to analyte i and orthogonal to all other factors in $\mathbf{S}_{-i}$, and S/C ~ | NAS|
- Adding clutter tends to add something parallel to $\mathbf{s}_i$ thus lowering NAS
  - Increases estimation error

GLS

- GLS can be used for target detection, classification and quantification
  - need a model of the clutter and a spectrum of pure component(s)
  - no need for buckets of calibration samples
    - in some cases these can't be acquired
  - a.k.a. matched filter and Aitken estimator
    - Turin, George L., "An Introduction to Matched Filters." *IRE Transactions on Information Theory*, **6**(3) 1960: 311-329. (this is in a special issue on matched filters) and is used extensively in the remote sensing community [e.g., Burr T, Hengartner N (2006) Sensors 6:1721-1750] and has also been referred to as an "adaptive matched filter" to highlight the fact that the clutter covariance can be easily modified resulting in a new filter.
    - Aitken, A., "On Least Squares and Linear Combinations of Observations", *Proceedings of the Royal Society of Edinburgh*, 1935, **55**, 42-48
    - E.g., 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.

GLS Weighting for ILS and PCA

- GLS weighting can be applied to inverse least squares models (e.g., PLS) and PCA

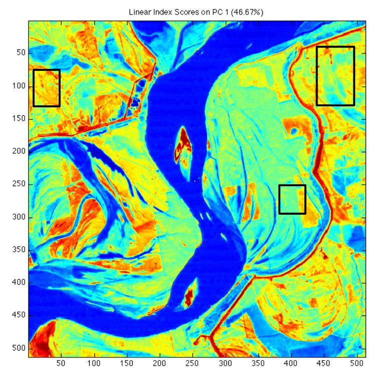
$\mathbf{Xb} = \mathbf{c}$ inverse least squares model

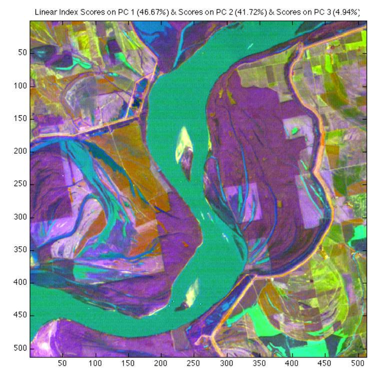
$\mathbf{X}_w = \mathbf{X}\mathbf{W}_c^{-1/2}$ weighting of **X** can be considered a generalization of autoscaling and is a pre-whitening step

can also be applied to standardization where the clutter covariance is the difference matrix between instruments

H. Martens, M. Høy, B.M. Wise, R. Bro and P.B. Brockhoff, "Pre-whitening of data by covariance-weighted pre-processing," *J. Chemo.*, **17**(3), 153-165 (2003).

Landsat Image of Mississippi

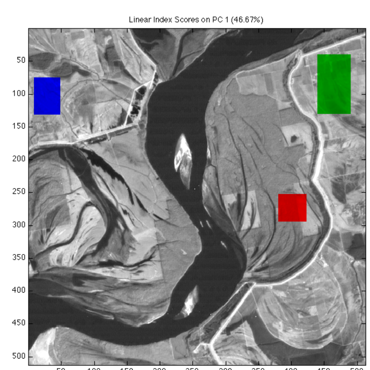


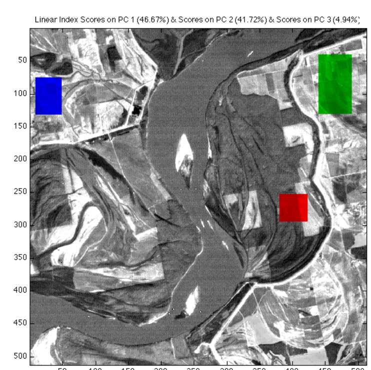


Scores on PC 1

Scores on First 3 PCs

Select Classes with Clutter to Down-weight

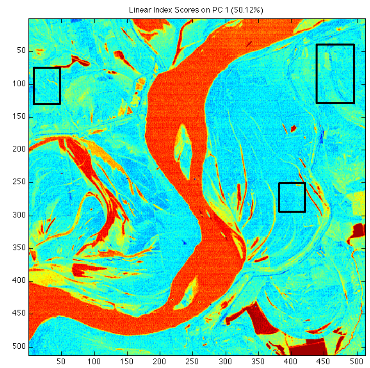


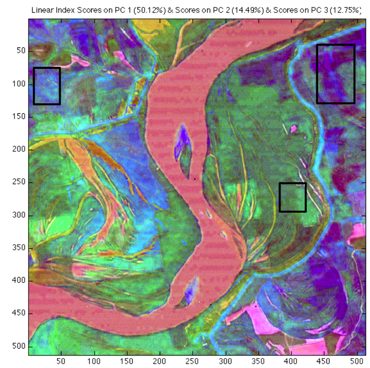


Scores on PC 1

Scores on First 3 PCs

PCA after GLS-Weighting

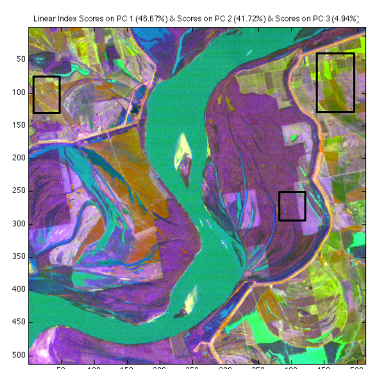


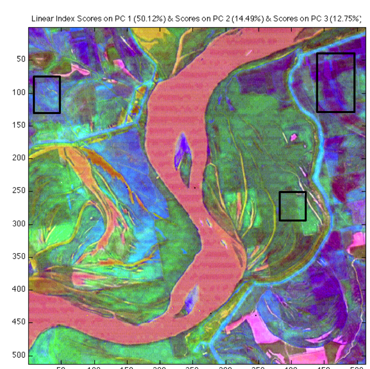


Scores on PC 1

Scores on First 3 PCs

PCA With and Without GLS-Weighting





Without GLS-Weighting

With GLS-Weighting

Comments on Filters

- Savitsky-Golay
  - For derivatives OR smoothing (noise reduction)
- Fourier
  - Remove high-frequency (noise) or low-frequency (baseline) components
  - Typically- NOT "windowed"
    - Position (wavelength) information not considered
- Wavelets
  - Extracting information by BOTH frequency and position
    - Allows BOTH feature selection and pre-processing!
  - filters that are based on window-size (scale)
    - orthogonal and oblique basis functions can be used
- Statistics w/in windows
  - Mean, Median, Max, Min

`SPATIAL_FILTER,LINE_FILTER, BOX_FILTER`

Multivariate Image Regression and Quantitative Analyses

- Inverse Least Squares models (Partial Least Squares – PLS)
- Classical Least Squares (CLS)
- Multivariate Curve Resolution (MCR)

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
    - **y = Xb**
    - predict a property
    - used for PLS-descriminant analysis

EIGENVECTOR RESEARCH INCORPORATED

Inverse Least Squares

- Inverse least squares (ILS) models assume that the model is of the form: $\mathbf{Xb} = \mathbf{y} + \mathbf{e}$

  where **y** (Mx1) is a property to be predicted,

  X (MxN_x) is the measured response,

  e (Mx1) is an error vector, and

  b (N_xx1) is a vector of coefficients

- It is possible to estimate **b** from $\mathbf{b} = \mathbf{X}^+\mathbf{y}$

  where $\mathbf{X}^+$ is the pseudo-inverse of **X**

EIGENVECTOR RESEARCH INCORPORATED

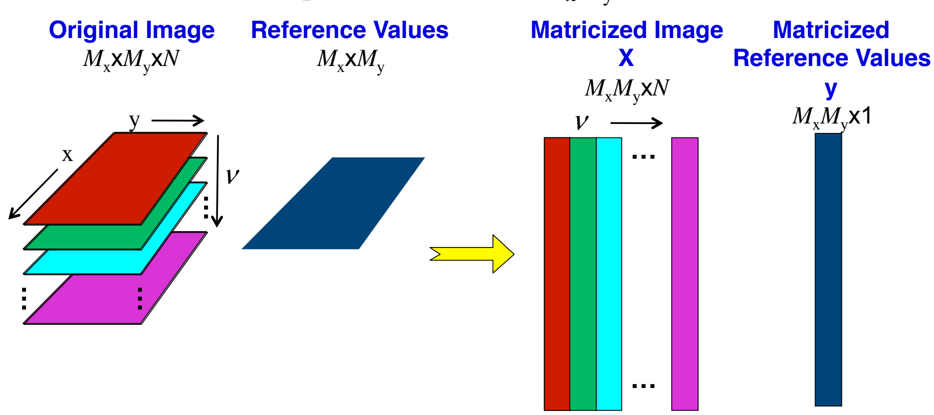
Advantage of ILS Methods

- ILS methods (including MLR, PCR, PLS, CR) don't require the concentration of all analytes, including interferents, be known …
- …however, interferents must vary in the calibration data set for the the ILS regression model to be robust against them
  - clutter factors must vary in the calibration data – it's best if they vary such that they are orthogonal to the target of interest
- Disadvantage is that reference values must be available in a representative number of pixels

EIGENVECTOR RESEARCH INCORPORATED

Unfolding ILS

- The image is M_xxM_yxN and it is reshaped by matricizing such that each pixel is a row in a M_xM_yxN matrix



EIGENVECTOR RESEARCH INCORPORATED

Estimation of b

- There are many ways to obtain a pseudo-inverse
- Multiple linear regression (MLR)[1] $\mathbf{X}^{+} = \left(\mathbf{X}^{T}\mathbf{X}\right)^{-1}\mathbf{X}^{T}$
  - problems with rank deficiency and ill conditioning
- Principal components regression (PCR)[2] $\mathbf{X}^{+} = \mathbf{P}_{K}\left(\mathbf{T}_{K}^{T}\mathbf{T}_{K}\right)^{-1}\mathbf{T}_{K}^{T}$
- Partial least squares (PLS)[2,3] $\mathbf{X}^{+} = \mathbf{W}_{\mathrm{k}}\left(\mathbf{P}_{\mathrm{k}}^{\mathrm{T}}\mathbf{W}_{\mathrm{k}}\right)^{1}\left(\mathbf{T}_{\mathrm{k}}^{\mathrm{T}}\mathbf{T}_{\mathrm{k}}\right)^{1}\mathbf{T}_{\mathrm{k}}^{\mathrm{T}}$
  - cross-validation used to select number of factors

[1]Draper, N. and Smith, H., “Applied Regression Analysis, Second Edition”, John Wiley & Sons, New York, NY (1981).
[2]Martens, H. and Næs, T., “Multivariate Calibration”, John Wiley & Sons, New York, NY (1989).
[3]M. Andersson, "A comparison of nine PLS1 algorithms," *J. Chemom.*, **23**(10), 518-529 (2009)

Model Performance Measures

- Average measures
  - Root mean square error of calibration (RMSEC)
    - measure of fit
  - RMSE cross-validation (RMSECV)
    - approximate measure of prediction
  - RMSE prediction (RMSEP)
    - measure of prediction
- Estimation error includes leverage

$$\mathrm{RMSEC} = \sqrt{\frac{\sum_{m=1}^{M}(y_m - \hat{y}_m)^2}{M}}$$

$$\mathrm{RMSECV}_K = \sqrt{\frac{\sum_{j=1}^{J}\sum_{m=1}^{M_j}(y_m - \hat{y}_m)^2}{M}} = \sqrt{\frac{\mathrm{PRESS}_K}{M}}$$

$$\mathrm{RMSEP} = \sqrt{\frac{\sum_{M=1}^{M_p}(y_m - \hat{y}_m)^2}{M_p}}$$

Faber, N.M. and Bro, R., Chemomem. and Intell. Syst., 61, 133-149 (2002).

For PCR and PLS: Number of PCs or LVs

- Choice is not always simple
- A few rules of thumb
  - sqrt(*M*) a good choice for number of splits
  - useful to do repeated CVs with different data ordering
    - want subsets to span the data space
  - be conservative, models are more often overfit than underfit
  - best choice is often not the global minimum PRESS
  - look for minimum of PRESS and work backwards if improvement is not at least 2%
  - RMSEC<RMSECV by more than ~20% indicates overfit
  - look at variance captured in **X** and **Y**. Is it significant with respect to what you know about the data?

Model Diagnostics

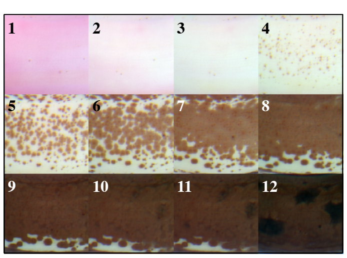
- Diagnostics useful for finding outliers/uniques
- **X**-block Q residual and T^2
- **X**-block leverage and studentized **Y**-block residuals

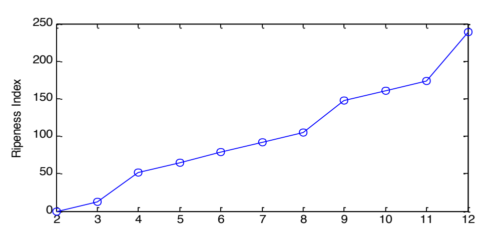
Unfold ILS Comments

- Easy, can use existing code with rearranged data
- Statistics reasonably well defined
- No second order advantage to be lost!
  - images are multi-mode but the spatial mode is not bi-linear

Banana Ripeness by PLS

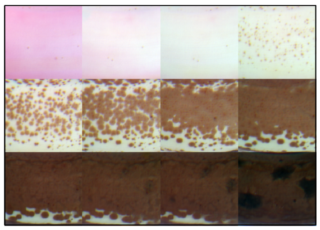
- Goal: Develop an automated (objective) method to assess banana ripeness
- X-Block RGB Images of Bananas at various stages of ripeness (Tiled)
- Y-Block Ripeness index for each tile



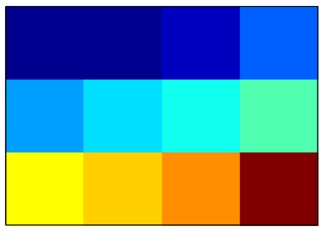


Data Courtesy Kim Esbensen
University of Ålborg, Denmark

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!)

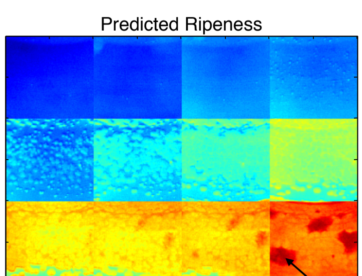
Y-block assumes a constant reference value for each image.

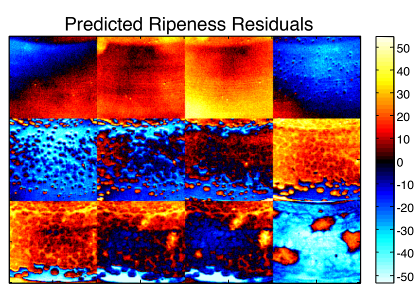
Unfold blocks before PLS

Note: Does not inherently take spatial correlation into account.

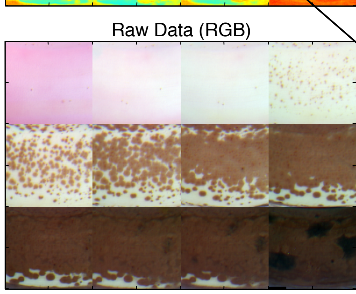
EIGENVECTOR RESEARCH INCORPORATED

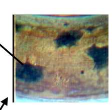
Banana Predictions





special colormap:
>>rkb



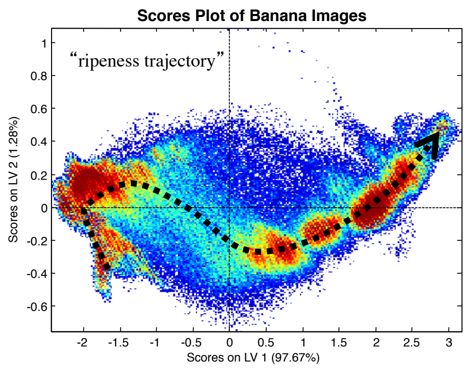


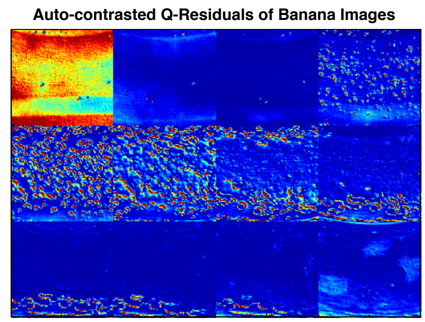
Autocontrasted image of last tile only

```
plotgui(bananas.imagedata(401:end,601:end,:))
%select   plot:slabs   and  view:autocontrast
```

EIGENVECTOR RESEARCH INCORPORATED

Banana Scores and Q Residuals





Classical Least Squares Models

- Classical Least Squares (CLS)
  - alternative to ILS models often used in imaging where reference values are unknown but target spectrum is known
  - extended mixture model, generalized least squares
    - accounting for clutter [unknown (but characterizable) interferences] in CLS
  - often used in spectroscopic applications and remote sensing

CLS Models

- Classical Least Squares (CLS)
  - also uses 'unfolded image'
  - requires spectrum of all chromophores
    - often cited as reason for ILS

$$\mathbf{X} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \qquad \hat{\mathbf{C}} = \mathbf{X}\mathbf{S}\left(\mathbf{S}^T\mathbf{S}\right)^{-1}$$

- Extended Mixture Model (ELS)

$$\mathbf{X} = [\mathbf{C} \quad \mathbf{T}][\mathbf{S} \quad \mathbf{P}]^T + \mathbf{E} \qquad \left[\hat{\mathbf{C}} \quad \hat{\mathbf{T}}\right] = \mathbf{X}[\mathbf{S} \quad \mathbf{P}]\left([\mathbf{S} \quad \mathbf{P}]^T[\mathbf{S} \quad \mathbf{P}]\right)^{-1}$$

  - where **P** is a sub-space that spance the systematic clutter variance
- Generalized Least Squares (GLS)

$$\mathbf{X} = \mathbf{C}\mathbf{S}^T + \mathbf{E} \qquad \hat{\mathbf{C}} = \mathbf{X}\mathbf{W}^{-1}\mathbf{S}\left(\mathbf{S}^T\mathbf{W}^{-1}\mathbf{S}\right)^{-1}$$

  - where **W** is the clutter covariance (might center also)
  - requires characterization of clutter
    - similar to requirement that interferences vary in ILS

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.

EIGENVECTOR RESEARCH INCORPORATED

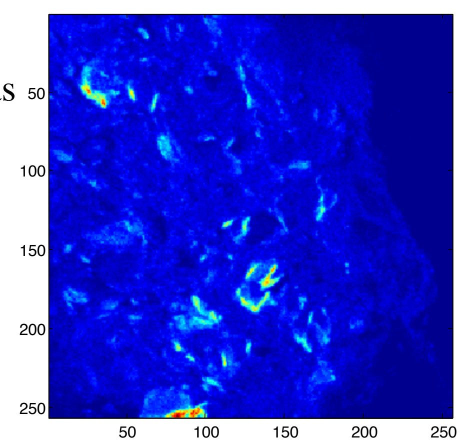
MCR Objective

- 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

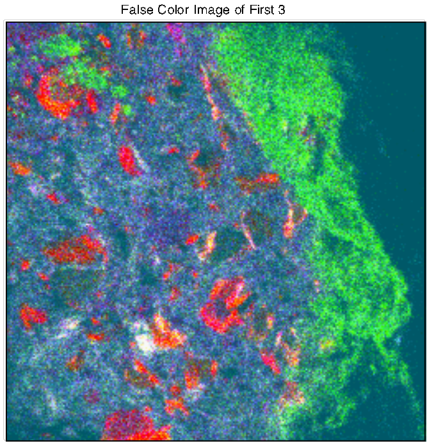
Imaging Mass Spec

- 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

PCA Score Image



Pretty picture, but loadings are very difficult to interpret!

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
- Showing Score Images – image was unfolded with each pixel as a separate sample then the scores are re-folded to form images

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).

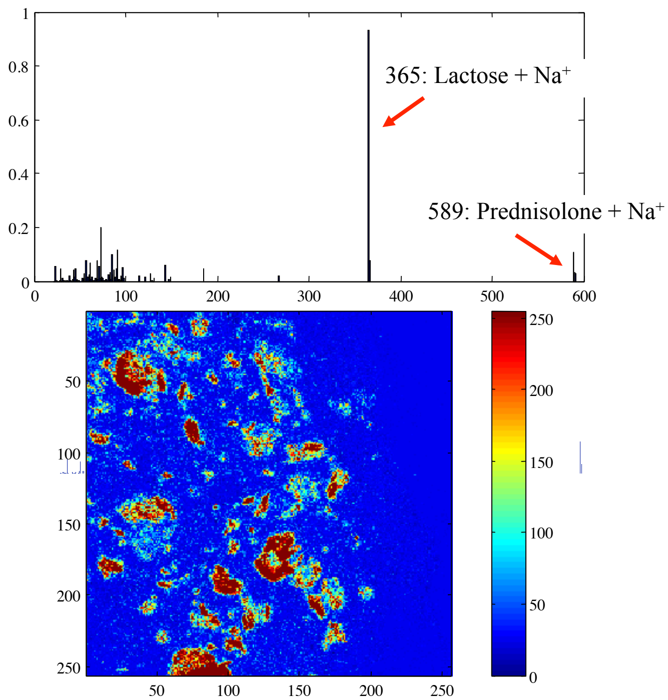
EIGENVECTOR RESEARCH INCORPORATED

Prednisolone:

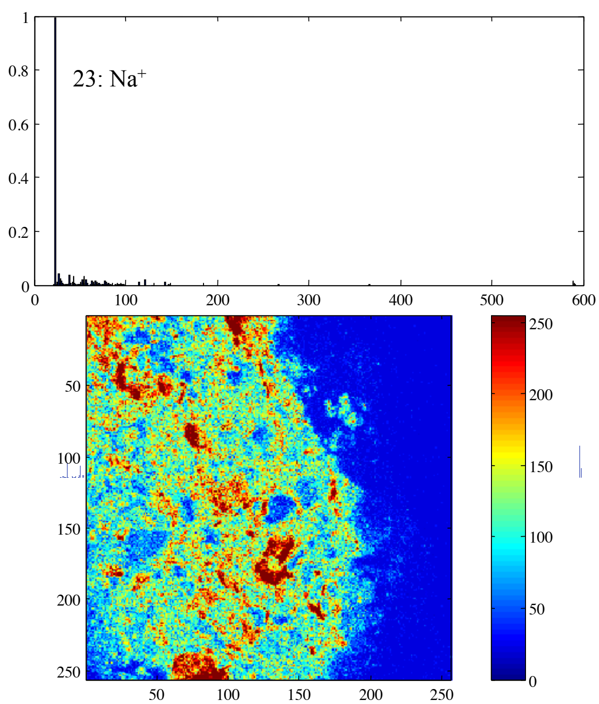
$C_{21}H_{31}NaO_9S$

Lactose:

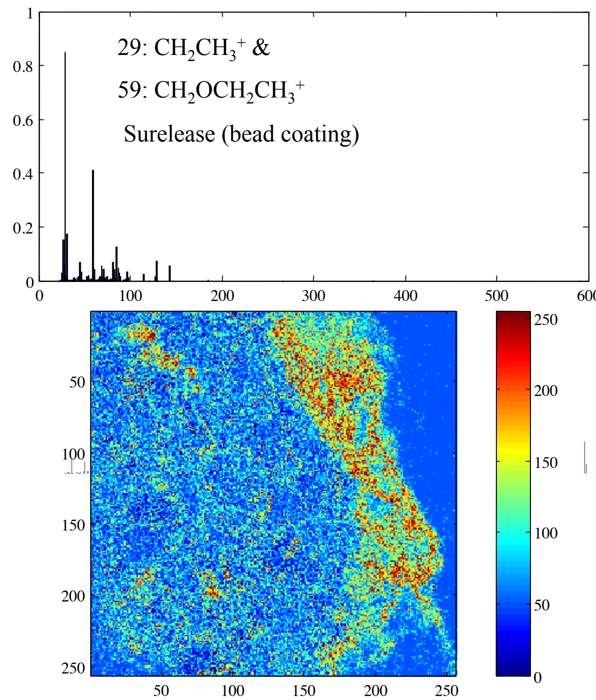
$C_{12}H_{22}O_{11}$



139



140



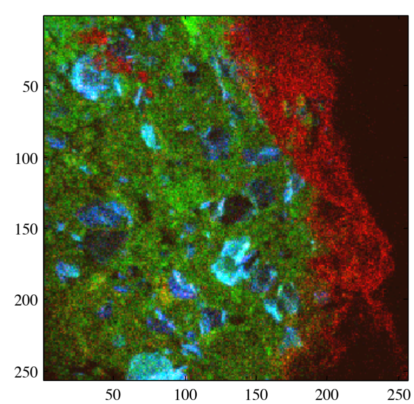
RGB "Chemical" Image

Red: Surelease (bead coating)

Green: Na

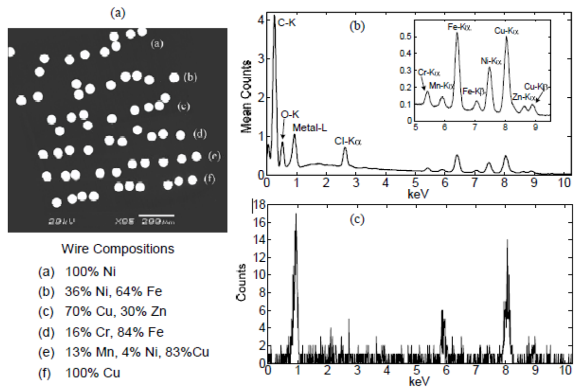
Blue: Prednisolone (drug)

only 3 of 6 factors extracted are shown



EIGENVECTOR RESEARCH INCORPORATED

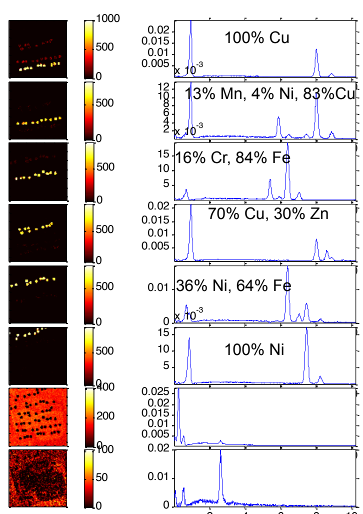
Energy Dispersive Spectrometry (EDS) Image of Wires



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

MCR Results on Wires



 Using Image Contrast (angle) constraint

EIGENVECTOR RESEARCH INCORPORATED

Example of Dealing w/ Clutter

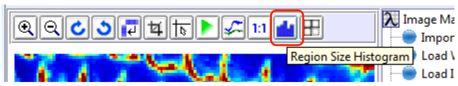
- MIA Example: Multivariate Curve Resolution (MCR)
    - Perform EMSC magnitude and slope correction (more later …)
        - reference is an estimate of the resin spectrum with robust fitting
        - allow glucose, lysine, $CaSO_4$ spectra to pass the filter
        - Gallagher, Blake, Gassman, *J. Chemometr.*, **19**(5-7), 271-281 (2005).
  - Step 2: Account for scratches using spatial constraints:
    - Scores from a PCA of region 2778 to 1790 cm^{-1} w/ 2^{nd} derivative preprocessing capture variability due to scratch features
    - Equality constraints on **C**: components 4 to 11 → the scratches
        - Soft equality Constraints on **S**: components 1 to 3
            - » Factor 1: resin, Factor 2: lysine (w/~ $CaSO_4$), Factor 3: glucose
  - Linear mixture model referred to as an extended mixture model

$$\mathbf{X} = \begin{bmatrix}\mathbf{C} & \mathbf{T}\end{bmatrix}\begin{bmatrix}\mathbf{S} & \mathbf{P}\end{bmatrix}^T + \mathbf{E}$$

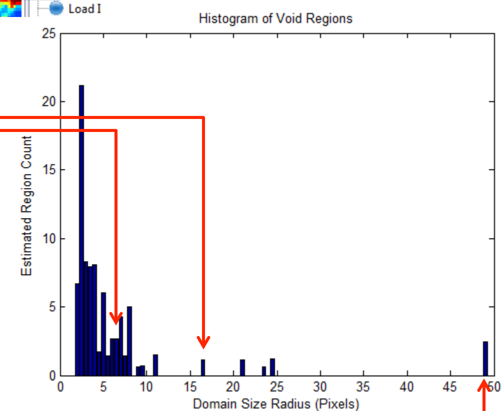
desired factors ($\mathbf{C}$, $\mathbf{S}$) — interferences ($\mathbf{T}$, $\mathbf{P}$)

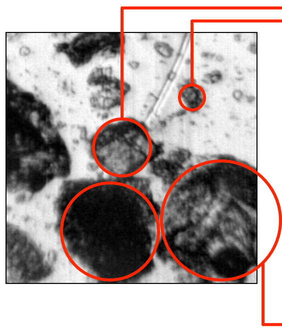
OTHER TOOLS

Radial Region Histograms



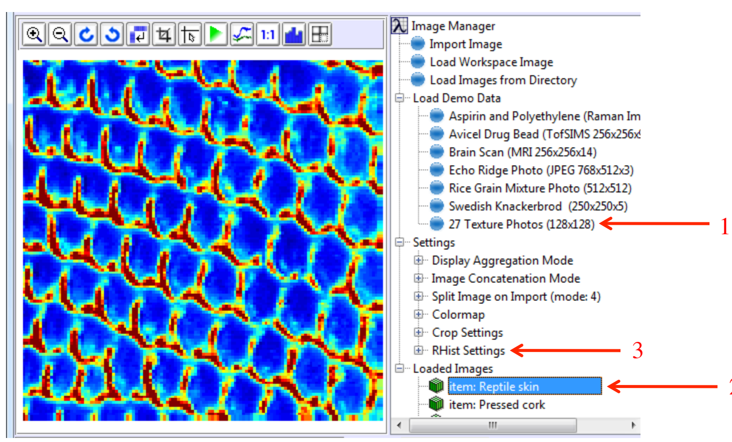
Fit circles of varying radius to
signal or voids of image.





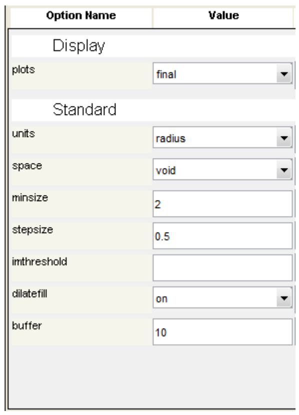
EIGENVECTOR RESEARCH INCORPORATED

Region Histogram Example: Reptile Skin



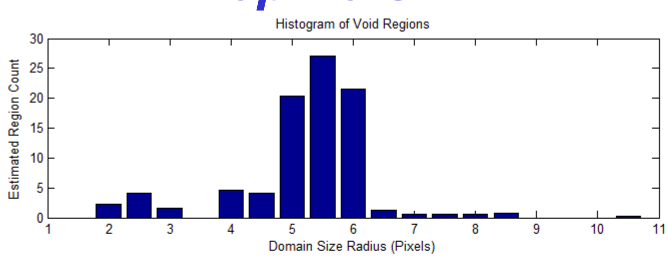
EIGENVECTOR RESEARCH INCORPORATED

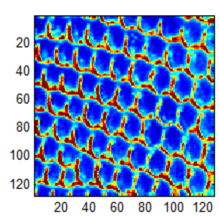
Region Histogram Settings

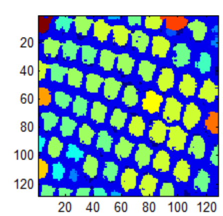


Plots –
Final (only histogram)
Detailed (includes filling map)
Units –
Radius, Diameter, Area
Space –
Signal: measure signal
Void: measure lack of signal
Dialatefill –
(On/Off) accommodates non-circular regions by adjusting circle to fill space

Region Histogram Example: Reptile Skin

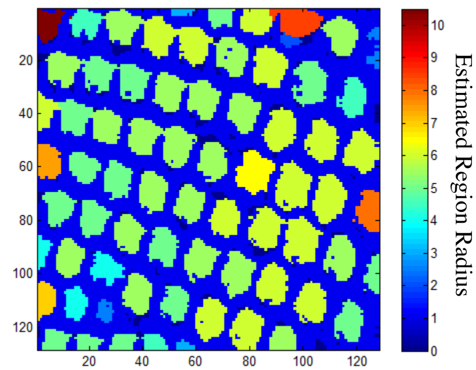






Hint: Right-click axes to spawn as separate figure.

Filled Region Map



EIGENVECTOR RESEARCH INCORPORATED