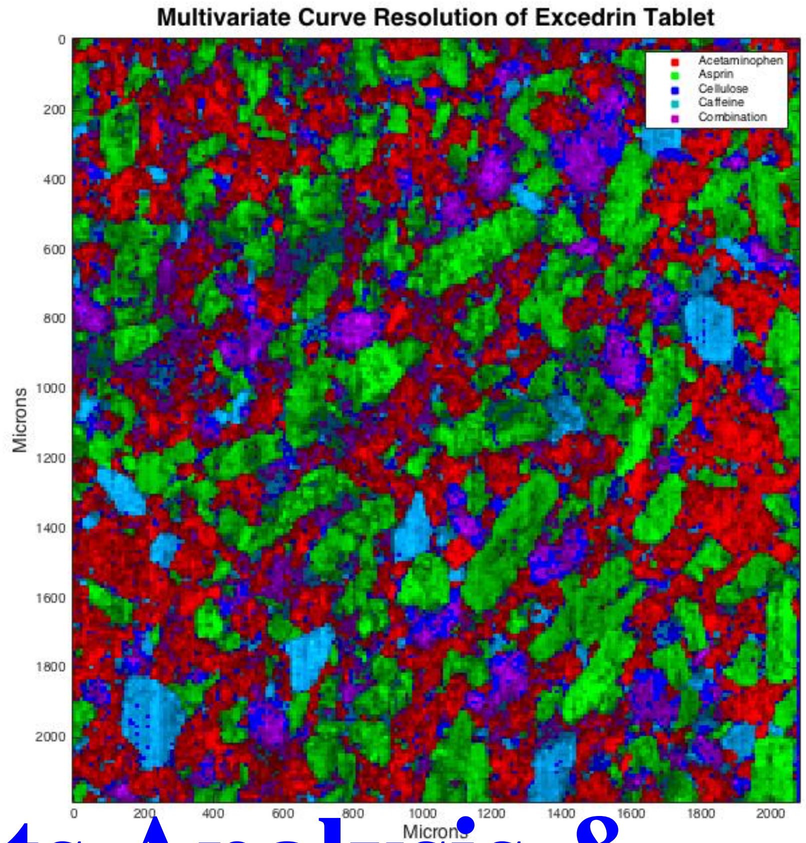



# Introduction to Principal Components Analysis & Related Methods on Multivariate/Hyperspectral Images

Barry M. Wise, Ph.D.

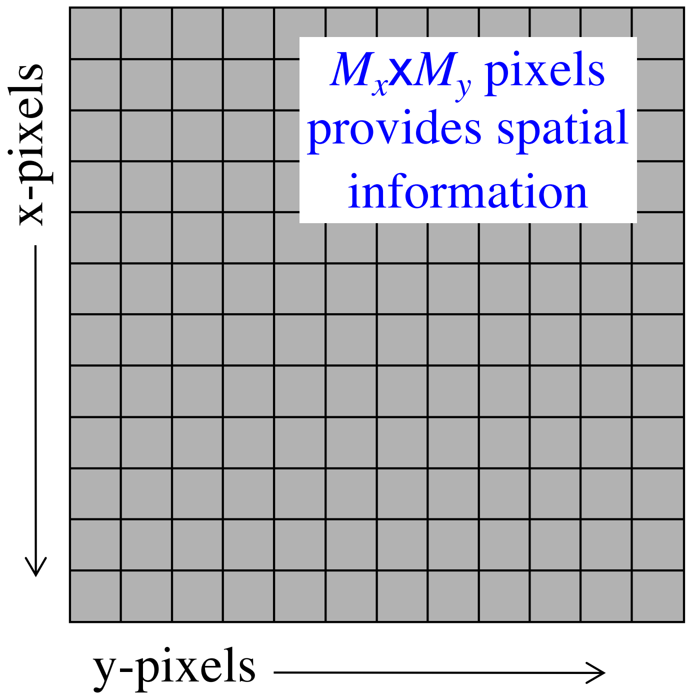
EIGENVECTOR RESEARCH INCORPORATED

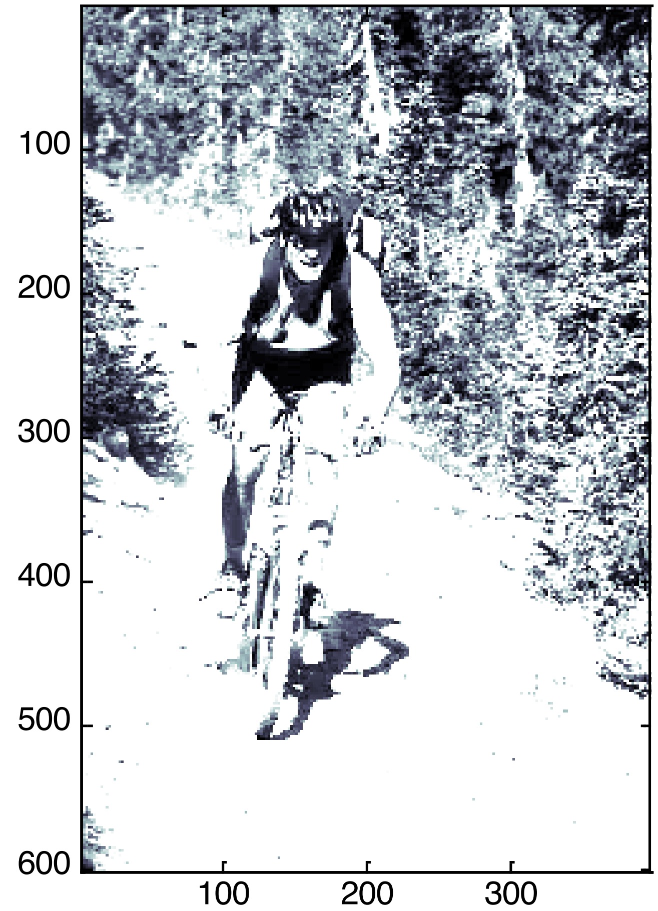
# *Outline*

- Intro to 3-way arrays and simple visualizations
- Principal Components Analysis (PCA)
- Multivariate Curve Resolution (MCR)
- Independent Components Analysis (ICA)
- Other methods: MAF, MDF, GLS, EPO
- Cluster Analysis (time permitting)
- Conclusions

# 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)

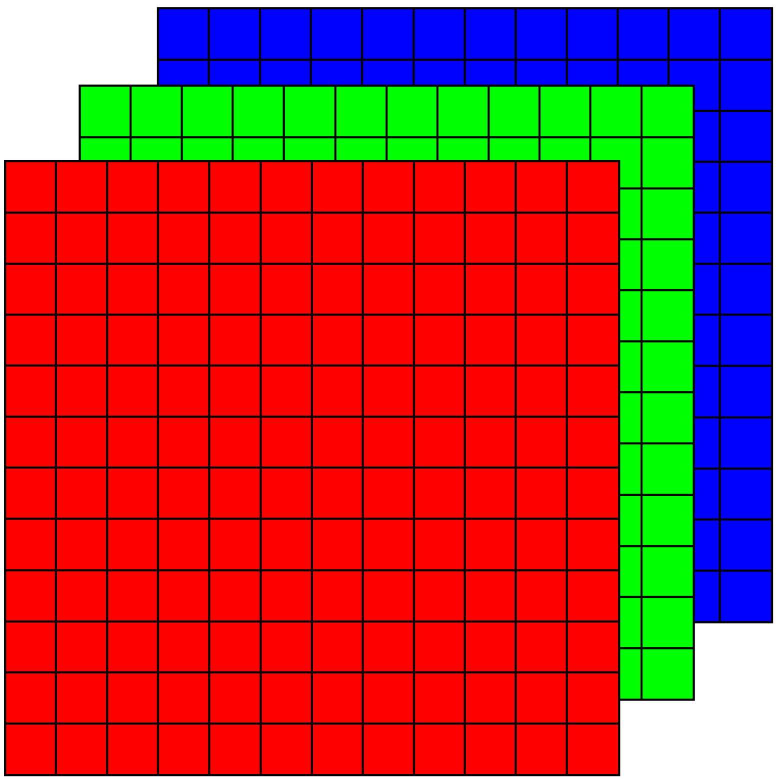


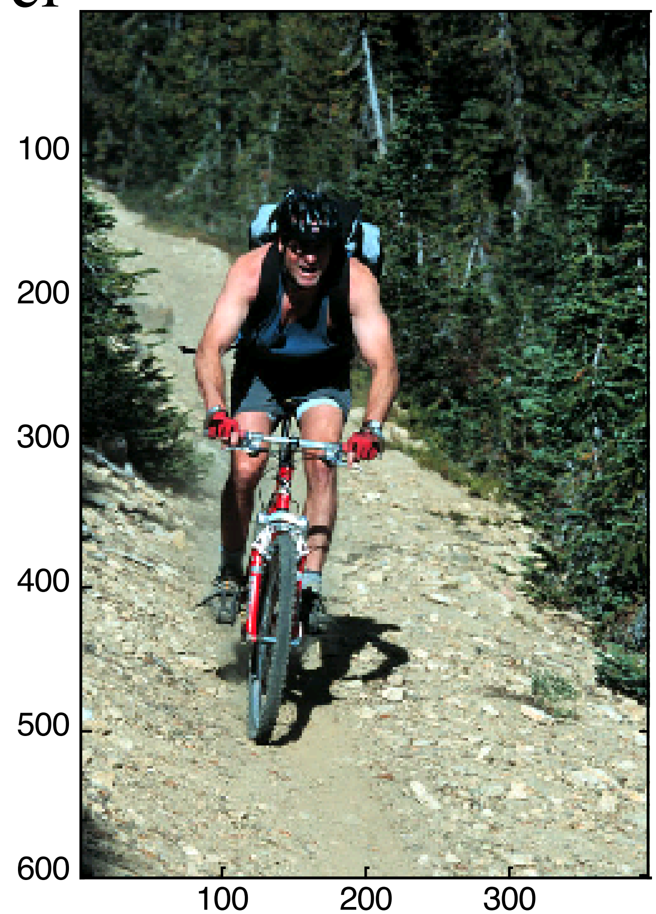




EIGENVECTOR RESEARCH INCORPORATED

# Multivariate Image
## (3 Variables)

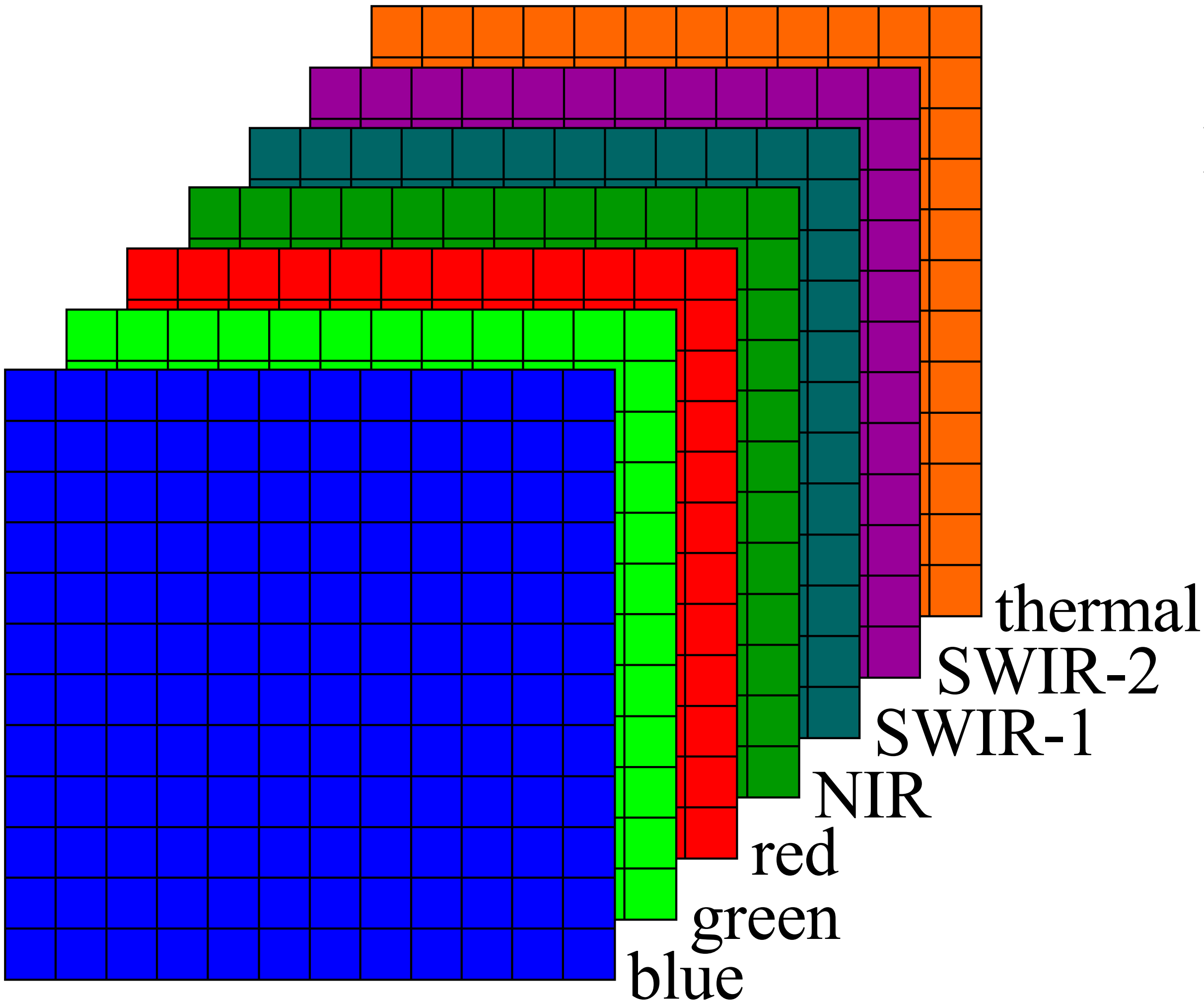
- Red/Green/Blue (RGB) (e.g. JPEG)
  - each layer defines color intensity level
  - much more information-rich





# Multivariate Image
## (4-10 Variables)

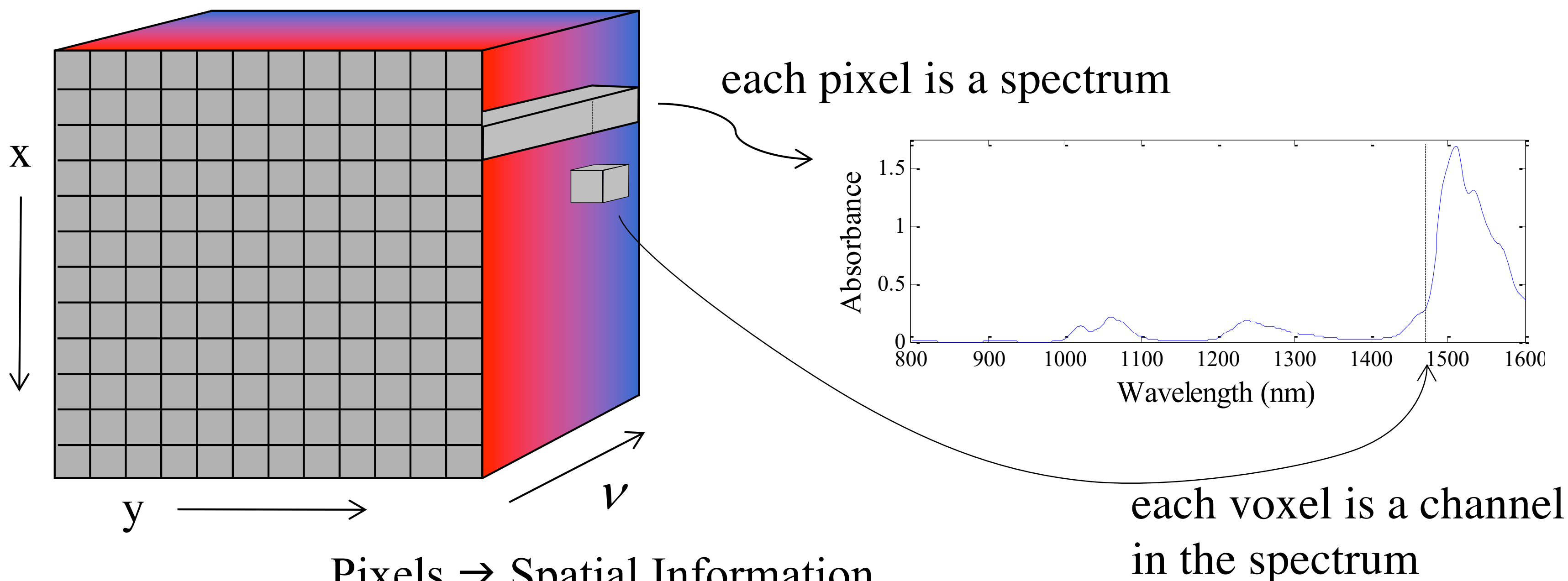
- Measure at several wavelengths (e.g., Landsat)




How should we display a seven variable image?

# 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

# 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
- 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 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, …

# Displaying a Multivariate Image

- Can choose any 3 variables (wavelengths) and display any image in RGB
- Doesn't choosing ignore potential information in the remaining variables?
- How could information be extracted from the multivariate image?
- Factor-based techniques—Principal Components Analysis
  - data reduction/compression
  - bring relevant information to surface
  - enhance S/N

Image of blue & NIR & thermal

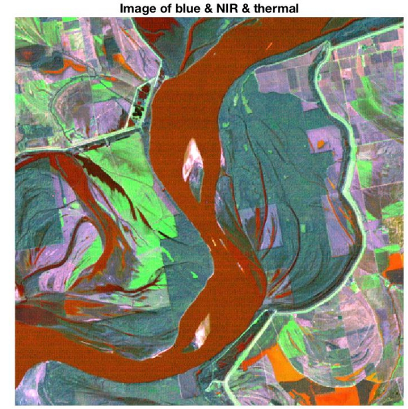
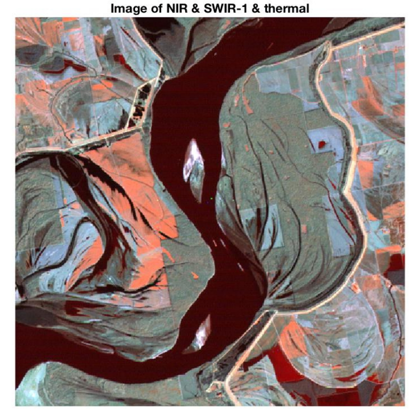


Image of NIR & SWIR-1 & thermal



EIGENVECTOR RESEARCH INCORPORATED

# Image Principal Components Analysis

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

# PCA Math

- 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**:

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

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

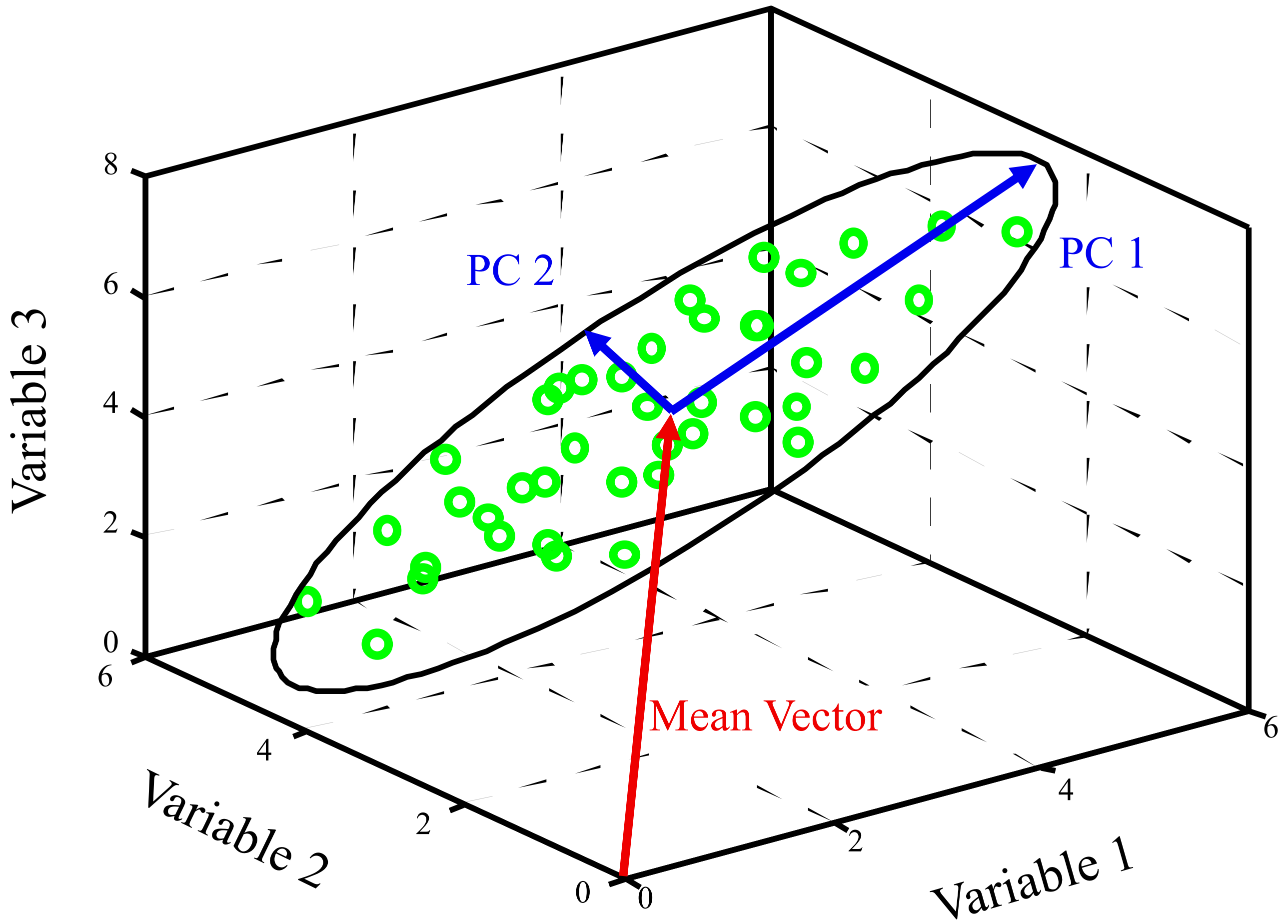
EIGENVECTOR RESEARCH INCORPORATED

# Properties of PCA

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

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

# PCA Graphically



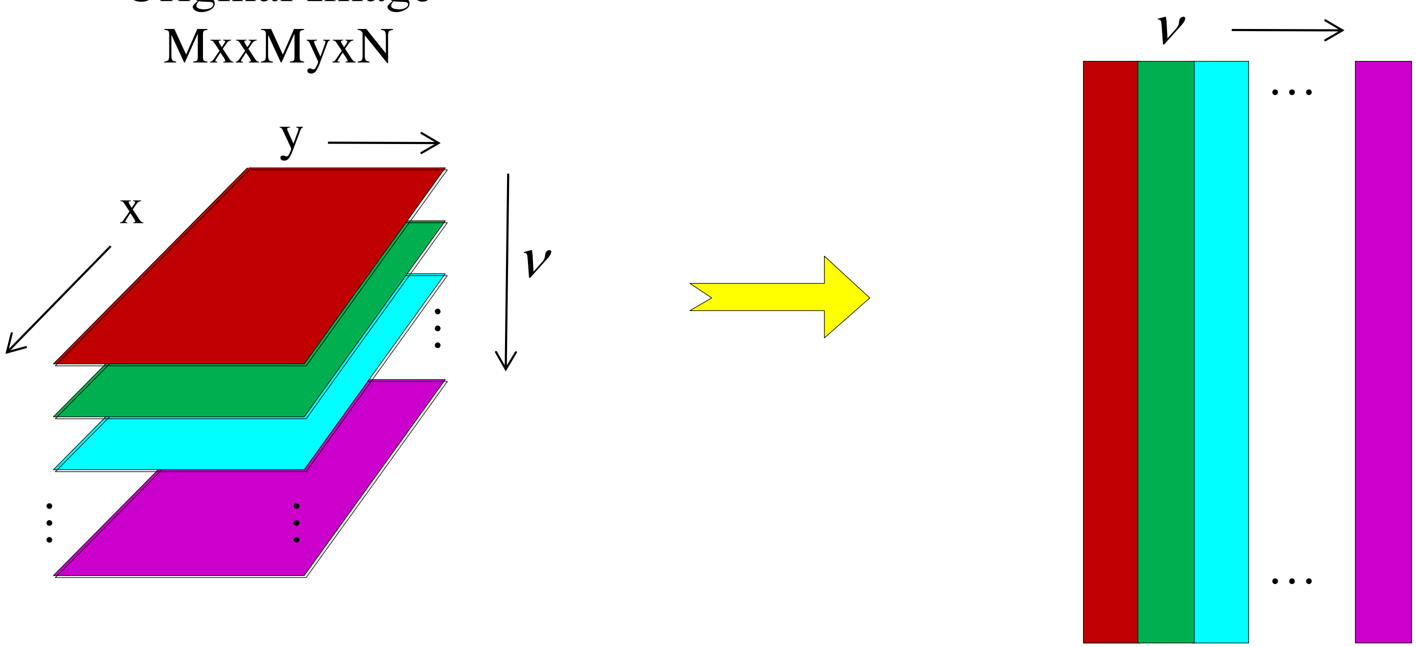

EIGENVECTOR RESEARCH INCORPORATED

# *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

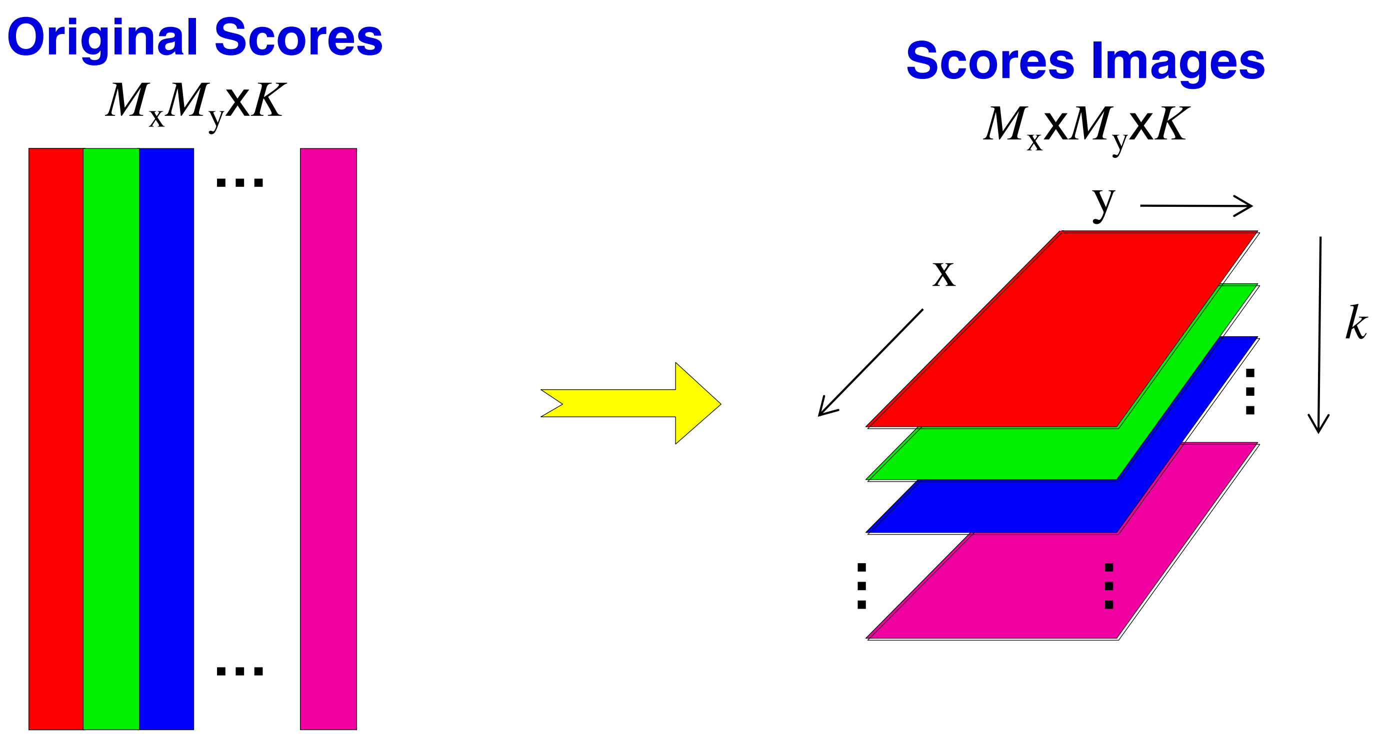
Original Image
MxxMyxN

Matricized Image
MxMyxN



# Reshape Scores To Images

- PCA gives scores **T** ($M$x$K$) which is reshaped to scores images ($M_x$x$M_y$x$K$)
  - each score vector is a $M_x$x$M_y$ scores image



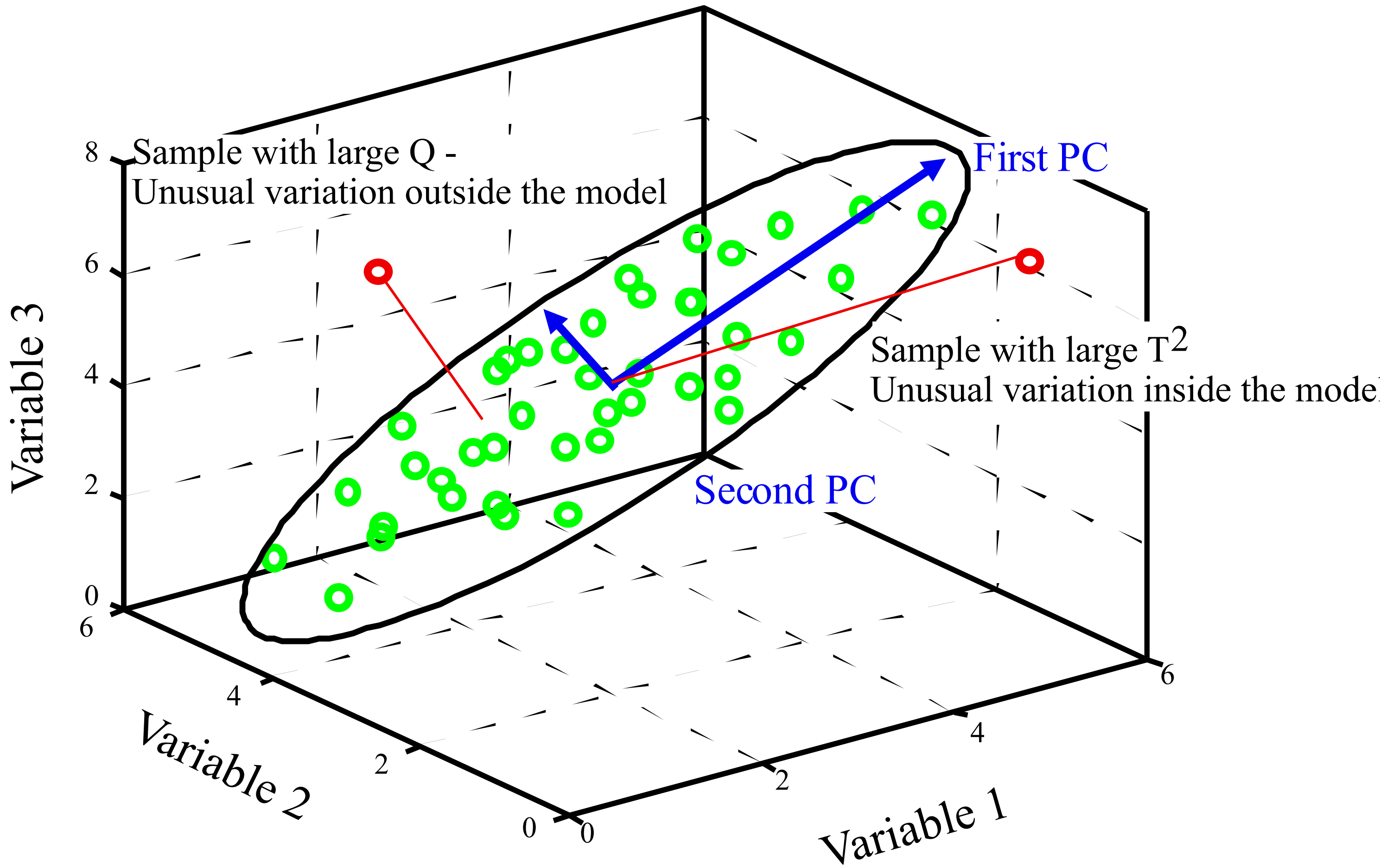

EIGENVECTOR RESEARCH INCORPORATED

# 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 $M_x$x$M_y$ 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

# Geometry of Q and $T^2$






EIGENVECTOR RESEARCH INCORPORATED

# 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 $\mathbf{X}$, $\mathbf{x}_m$, $m = 1, \ldots, M$

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

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

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

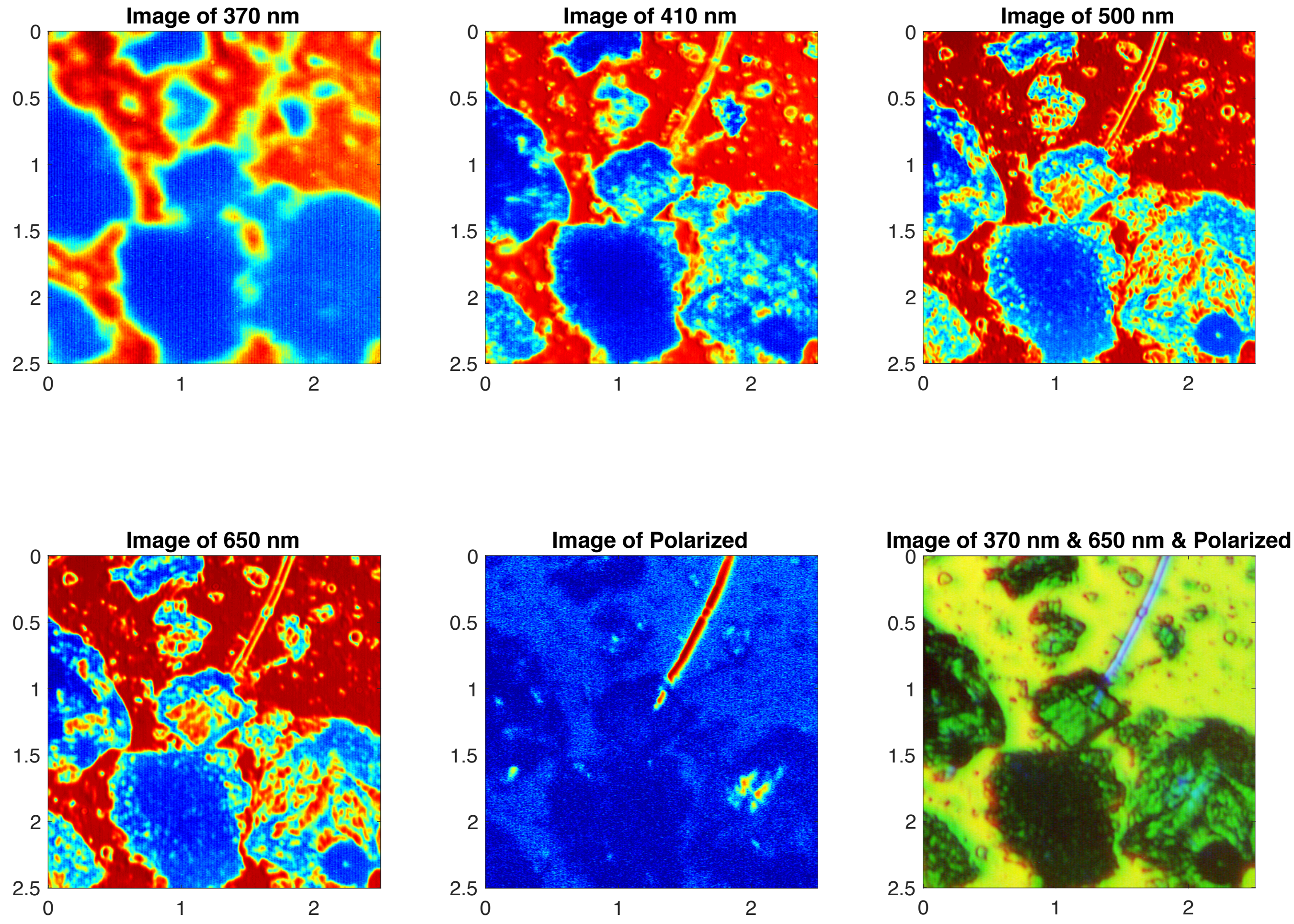
  - and also for individual columns:
    - scores, $\mathbf{t}_{mk}$
    - residuals $\mathbf{e}_{mk}$

EIGENVECTOR RESEARCH INCORPORATED

# Example: Bread

- Image of Swedish knäckebröd at 4 different wavelengths (370, 410, 500 & 650nm) plus visible light with a polarizing filter
- Image is 500 by 500 by 5
- Thanks to Paul Geladi for the data!

# Slices of the Bread



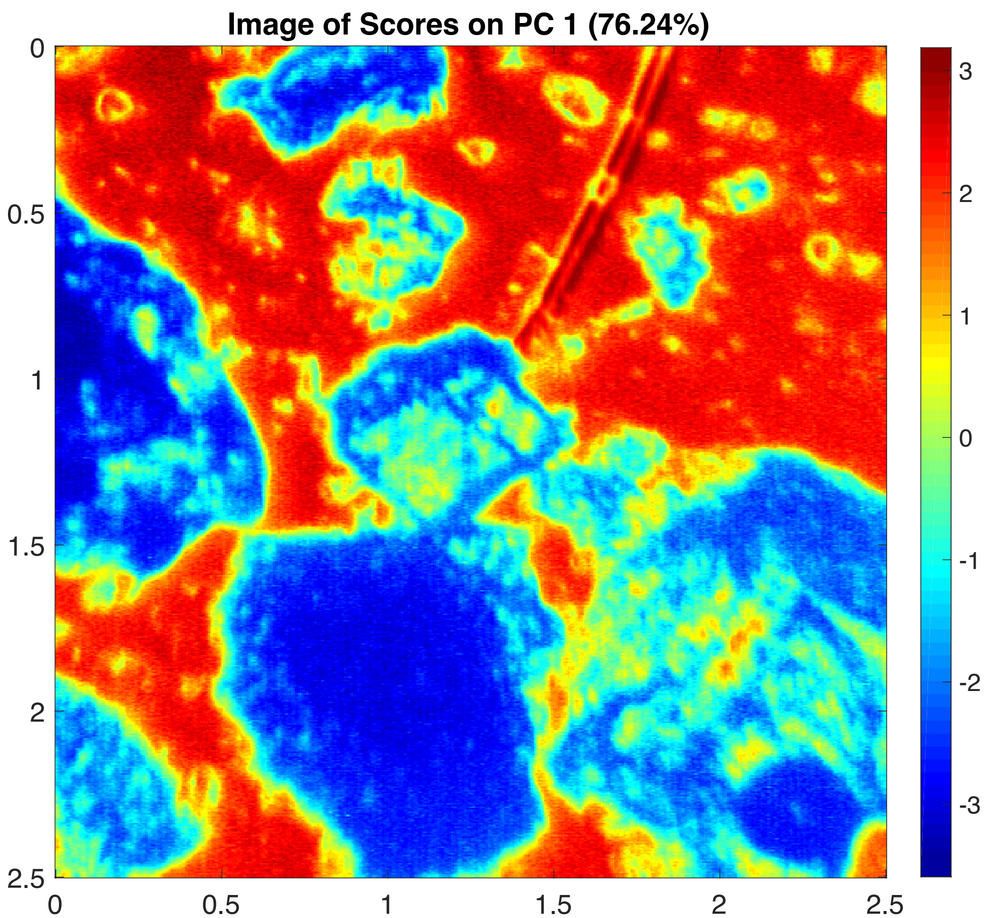

EIGENVECTOR RESEARCH INCORPORATED

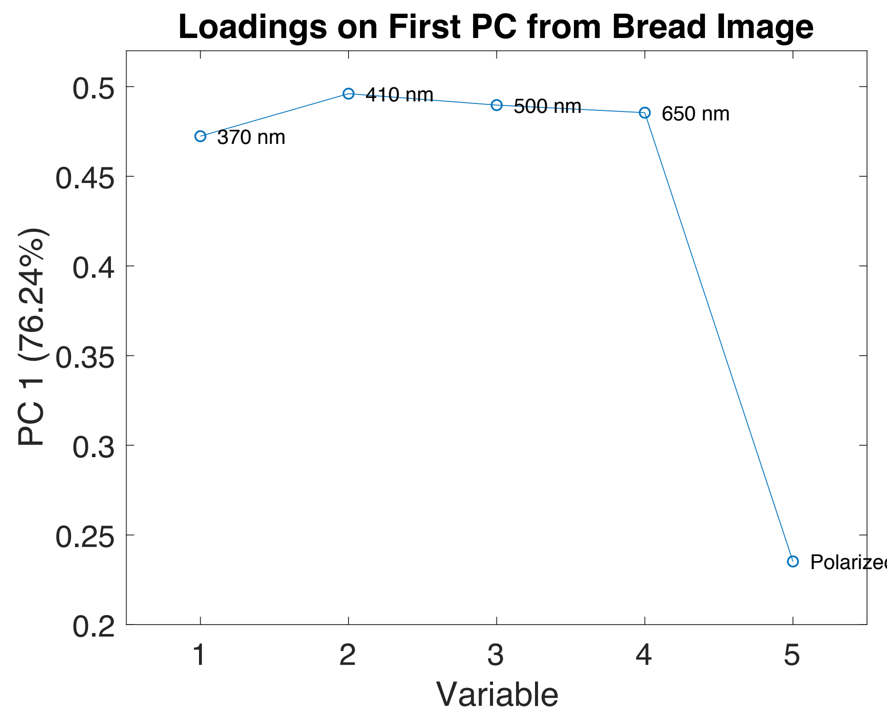
# Variance Captured Table

Percent Variance Captured by PCA Model

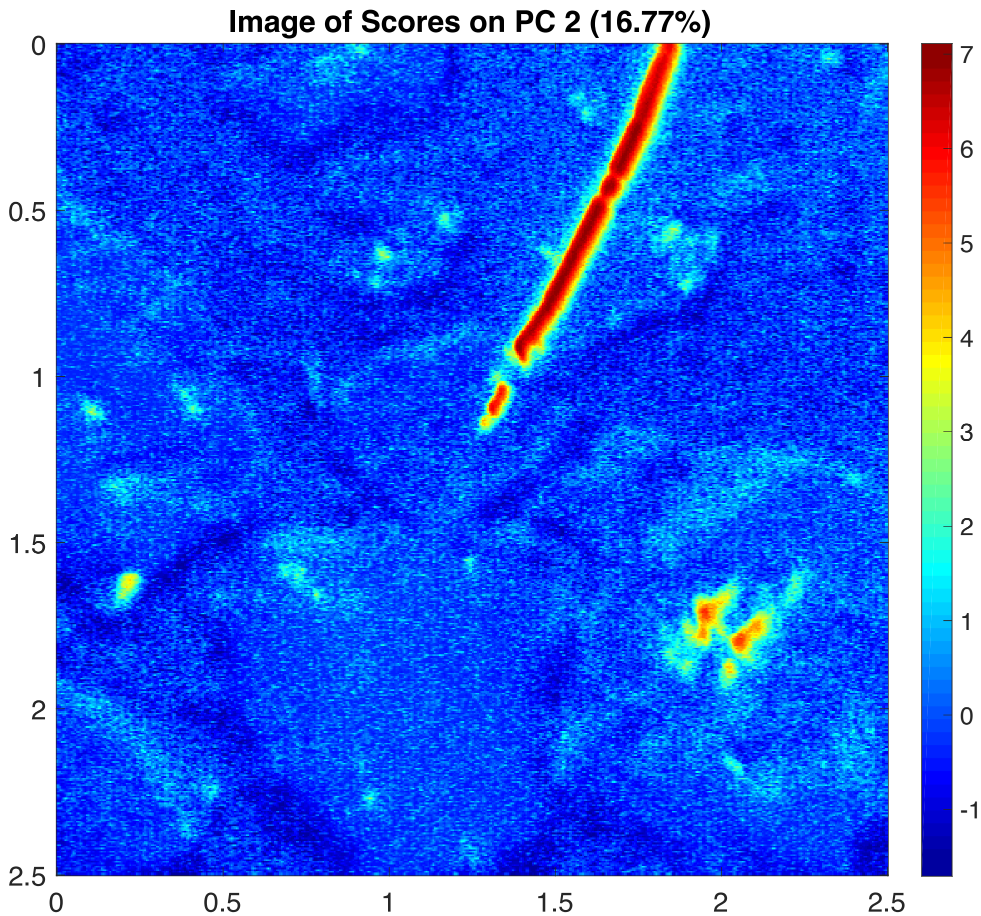
| Principal Component Number | Eigenvalue of Cov(X) | % Variance Captured This PC | % Variance Captured Total |
|---|---|---|---|
| 1 | 3.81e+00 | 76.24 | 76.24 |
| 2 | 8.38e-01 | 16.77 | 93.00 |
| 3 | 2.25e-01 | 4.51 | 97.51 |
| 4 | 7.98e-02 | 1.60 | 99.11 |
| 5 | 4.46e-02 | 0.89 | 100.00 |

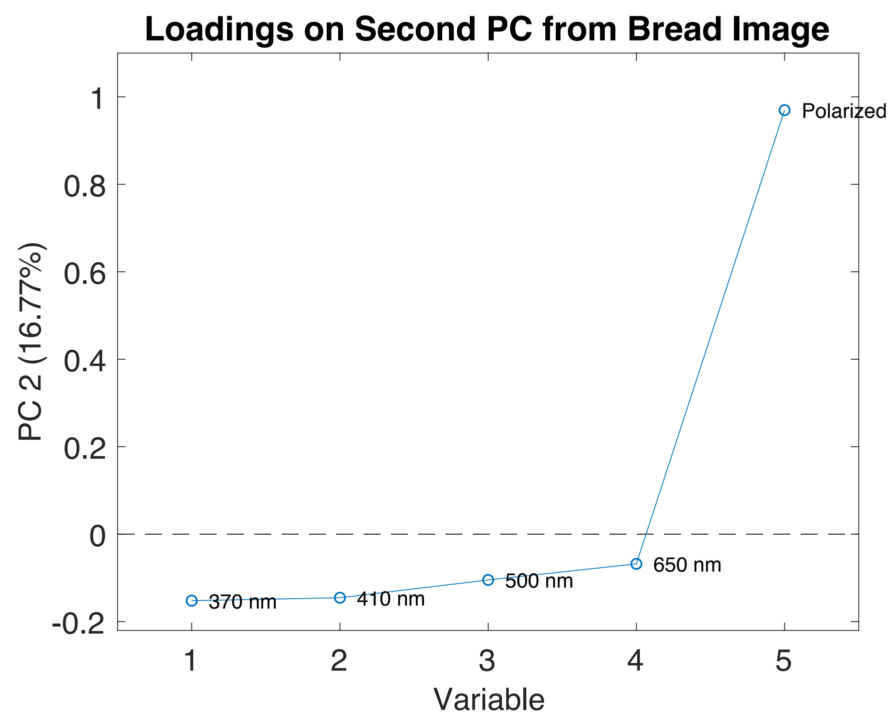
# First PC Scores & Loadings



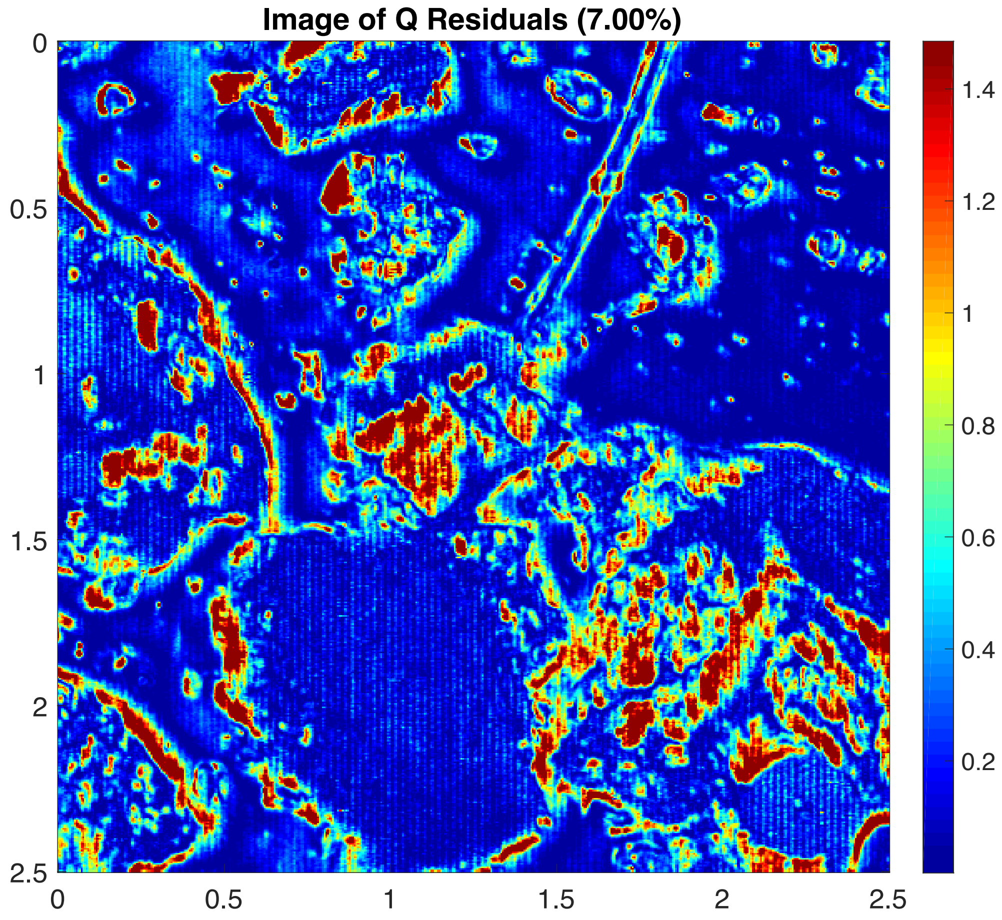



# Second PC Scores & Loadings



# Residual Image

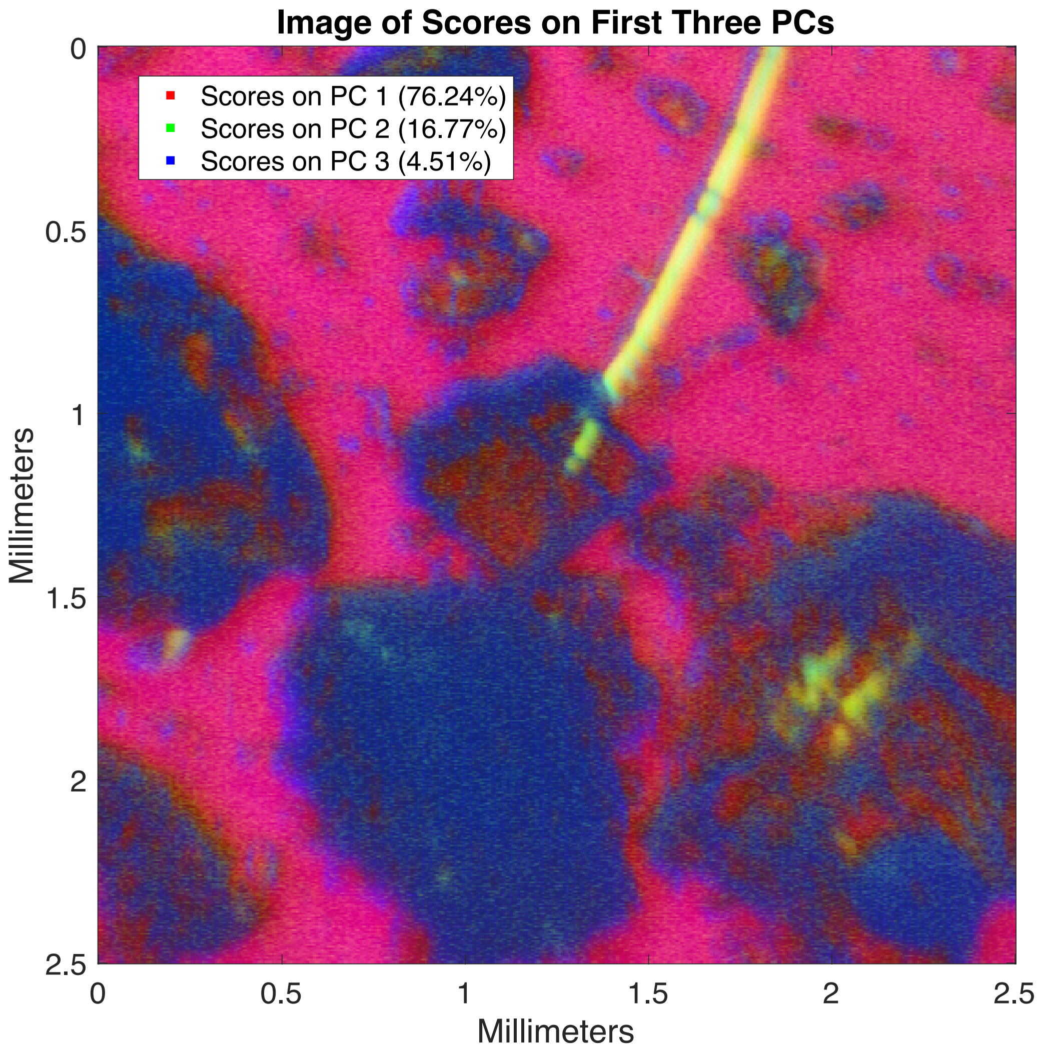



EIGENVECTOR RESEARCH INCORPORATED

# Creating False Color Images

- Color images are made up of three layers: red, green, blue
- Scores on the PCs can be used to define the intensity of each of the three layers
- Easy to do PC1 = red, PC2 = green, PC3 = blue

# False Color Image






EIGENVECTOR RESEARCH INCORPORATED

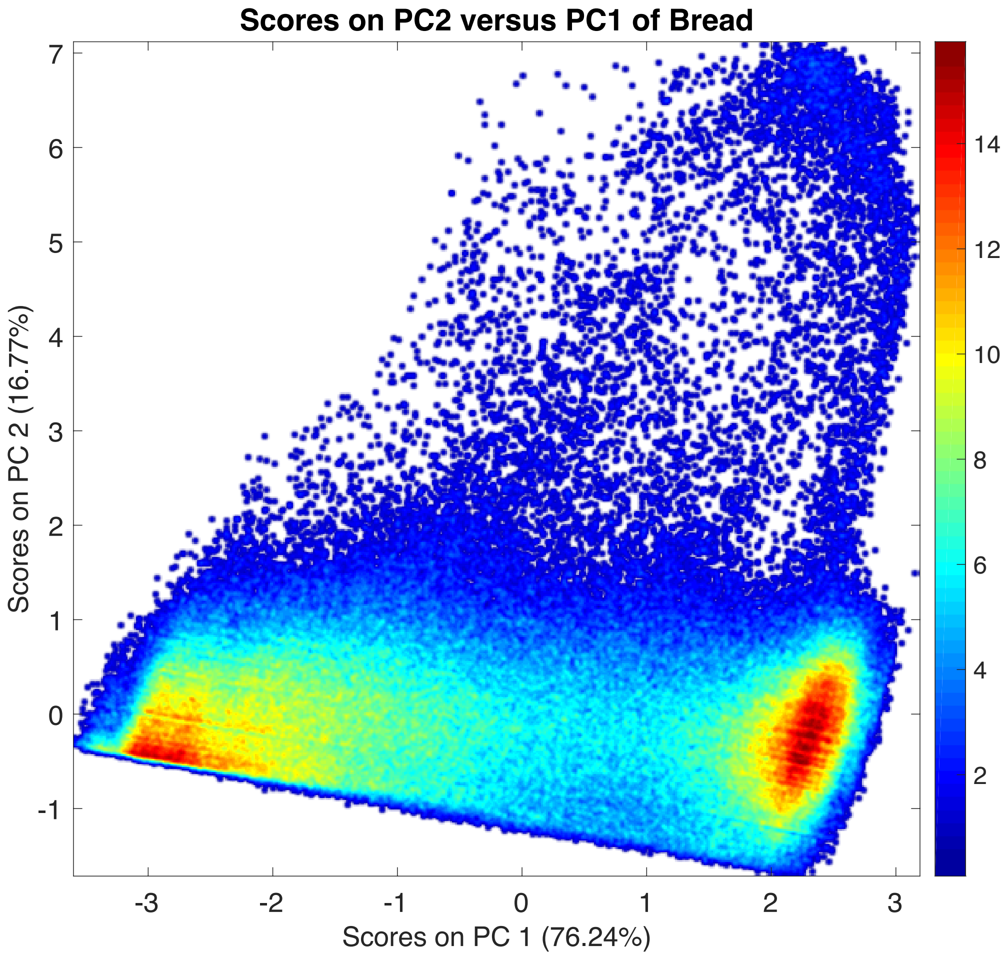
# Bivariate Scores Plots

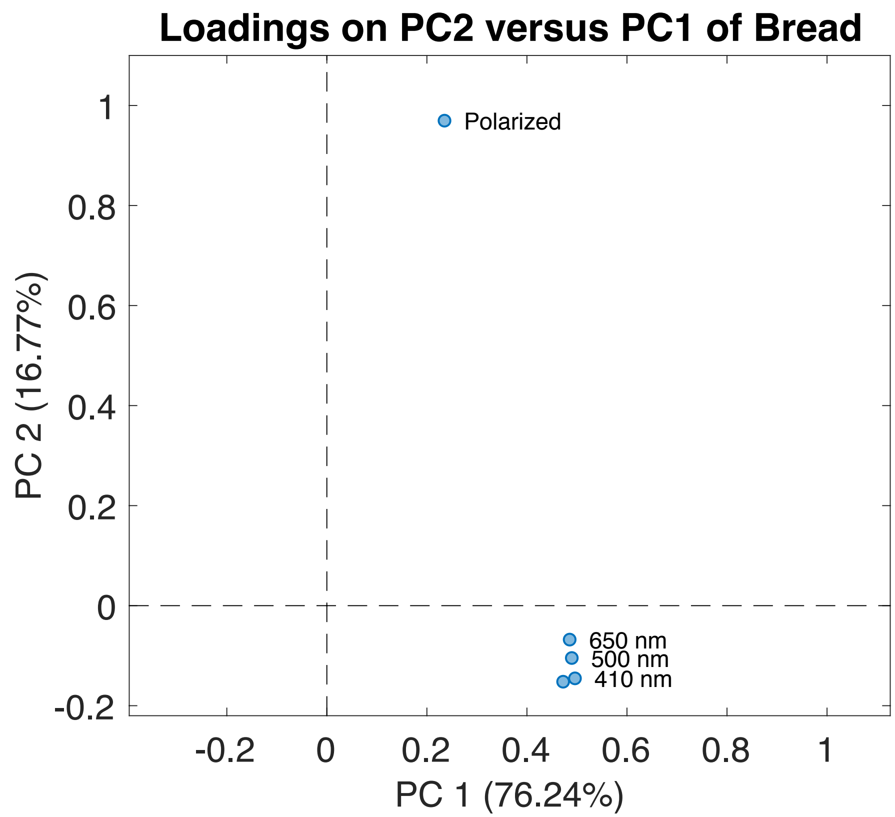
- Plot the scores for each pixel from one PC against the scores from another PC
- Problem: lots of points on the plot!
  - For this example: 500*500 = 250,000 points
  - Would look like single blob if plotted as one color
- Solution: score density plots
  - Calculate number of pixels with identical scores
  - Color code score plot according to number of pixels at each point
  - Can be useful to use log scale



EIGENVECTOR RESEARCH INCORPORATED

# Scores and Loadings



# Interpreting Scores and Loadings Plots

- Interpretation of scores and loading exactly the same as in conventional PCA
  - Look for clusters of pixels in scores plots
  - Use loadings to determine which original variables are responsible for differences in the scores
- Problem: what do scores clusters correspond to in the image plane?
- Solution: Linking

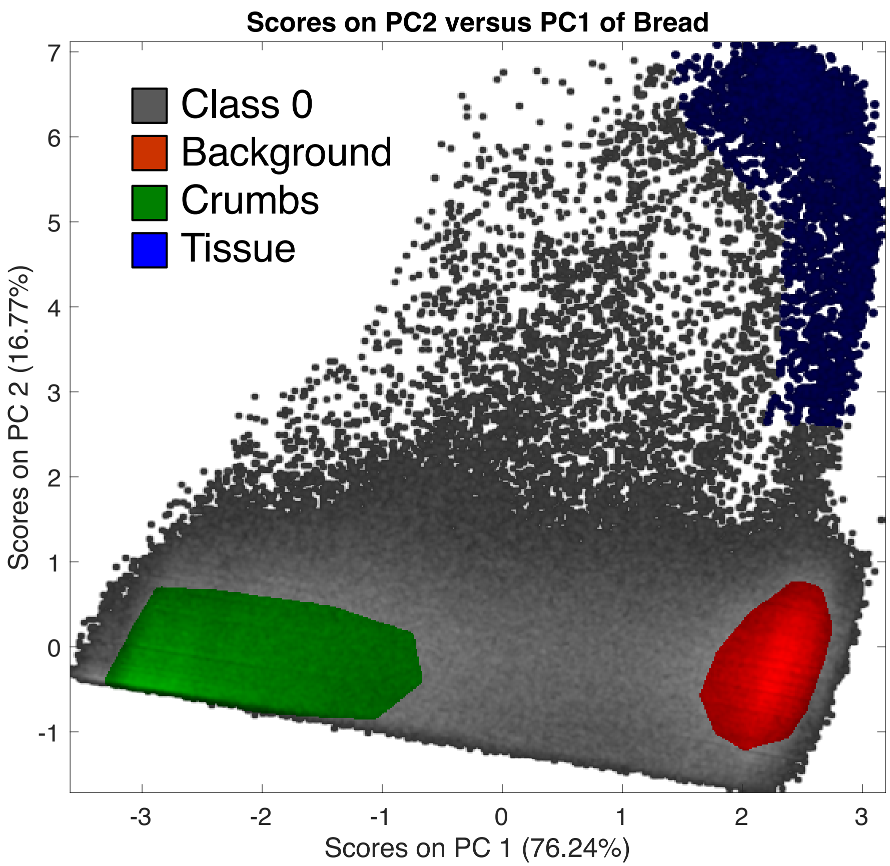
# Linking Scores and Image Plane Plots

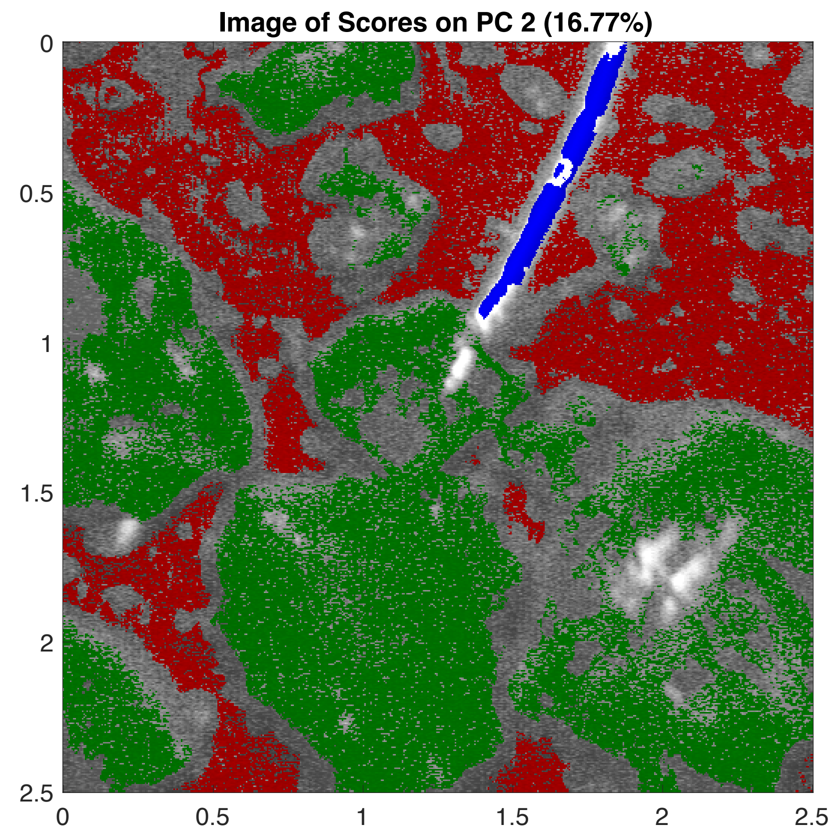
- Interpretation of image PCA models is easier when features in scores plots and the image plane are linked
- Pixels associated with scores in user defined polygons are highlighted in all plots
- Areas in image plane can also be linked to scores plots



EIGENVECTOR RESEARCH INCORPORATED

# Linked Scores Plots




EIGENVECTOR RESEARCH INCORPORATED

# Multivariate Curve Resolution

- With a minimum of *a priori* information, decompose a data matrix or image 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, …



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

# *Alternating Least Squares (ALS)*

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

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

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

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

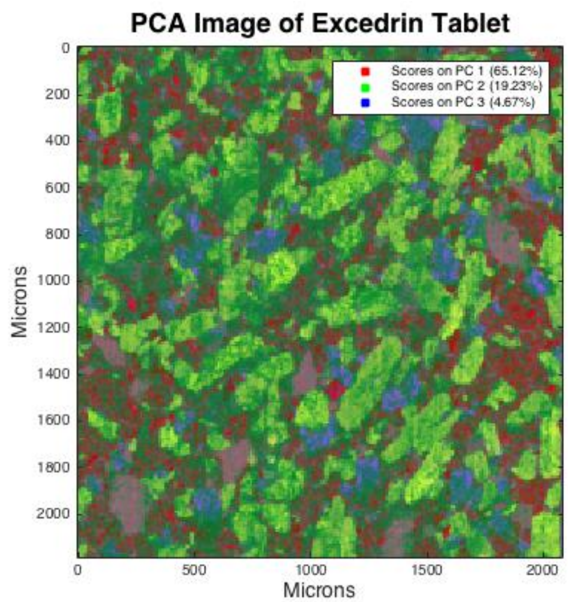
a.k.a. self-modeling curve resolution, self-modeling mixture analysis, end-member extraction



EIGENVECTOR RESEARCH INCORPORATED

# Example: MCR on Excedrin

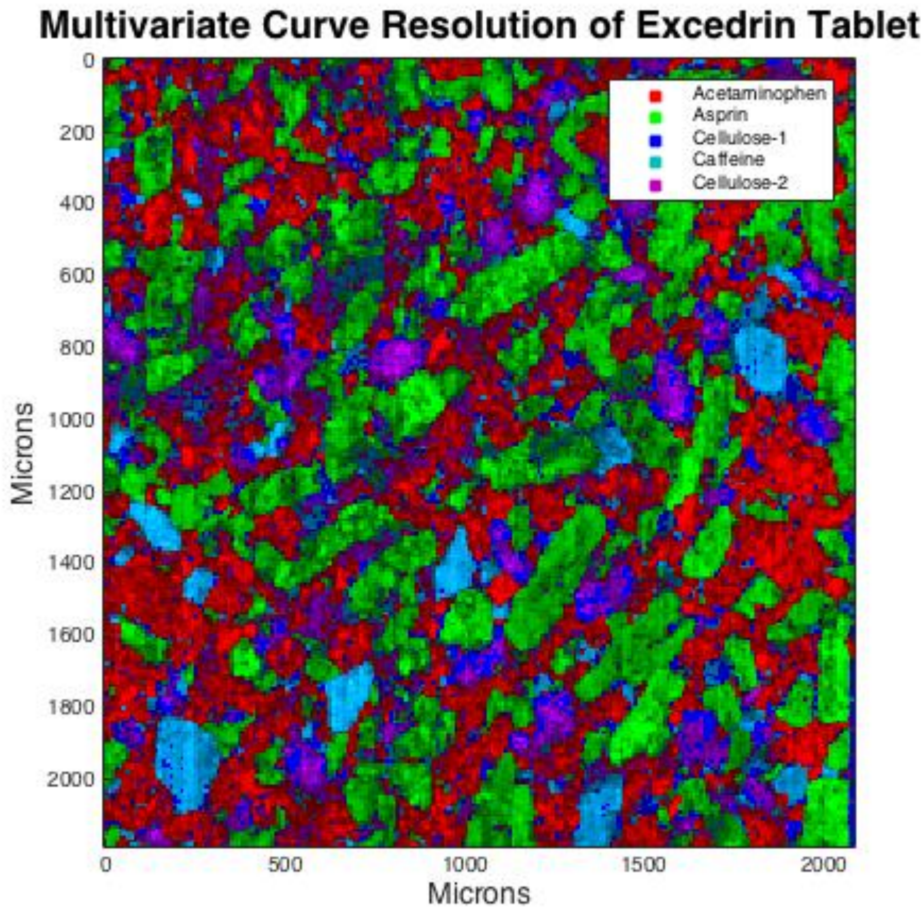
- Excedrin is a mixture of aspirin, acetaminophen, caffeine and microcrystalline cellulose
- Tablet imaged with tunable laser from 800 to 1800 $cm^{-1}$ over ~2mm
- Thanks to Agilent for data!

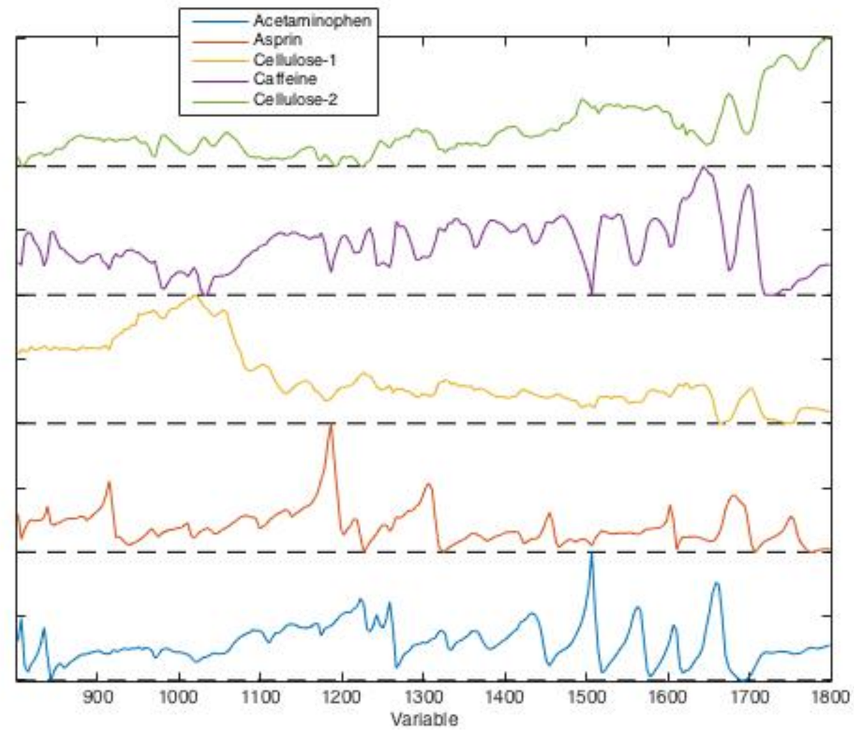




EIGENVECTOR RESEARCH INCORPORATED

# MCR on Excedrin Results






EIGENVECTOR RESEARCH INCORPORATED

# *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



EIGENVECTOR RESEARCH INCORPORATED

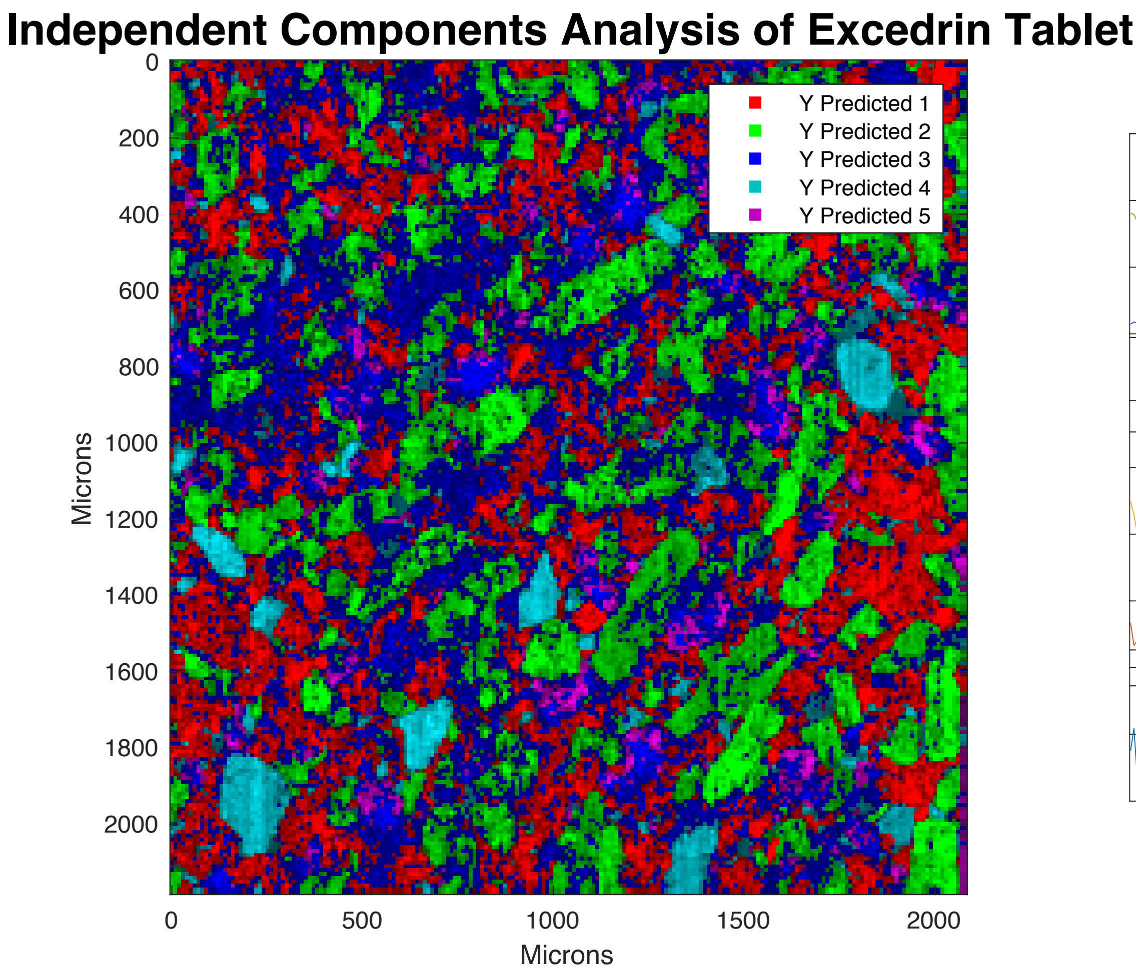
# Independent Components Analysis

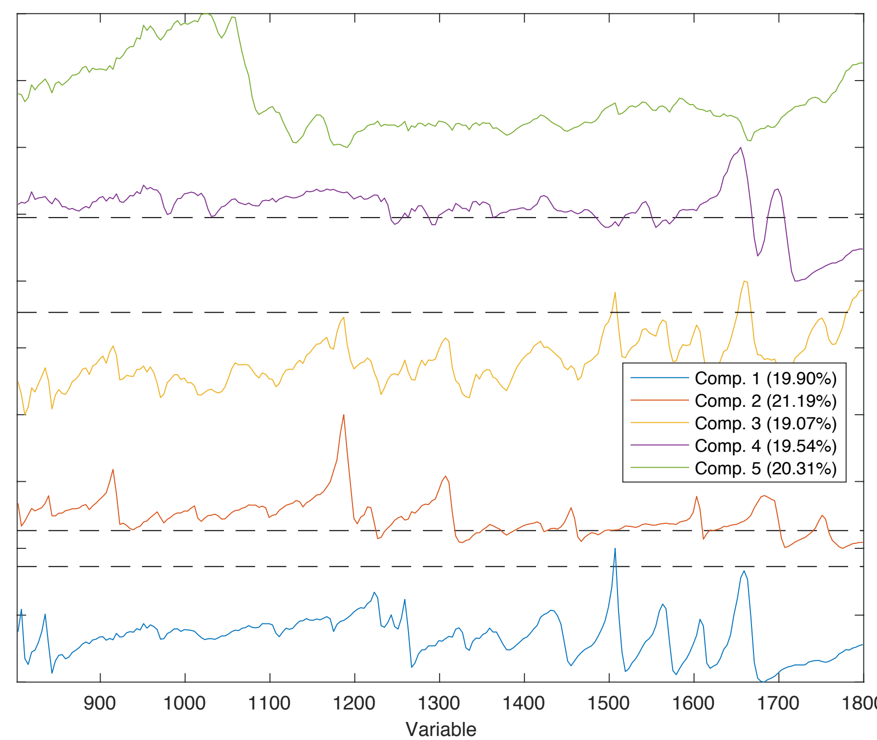
- Factor based method similar to PCA
- Central limit theorm says that sums of distributions tend towards Gaussian regardless of parent distributions
- ICA factors are computed to be as non-Gaussian as possible by maximizing the excess Kurtosis of the scores

$$k_{excess} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{\frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3$$

EIGENVECTOR RESEARCH INCORPORATED

# ICA on Excedrin Results

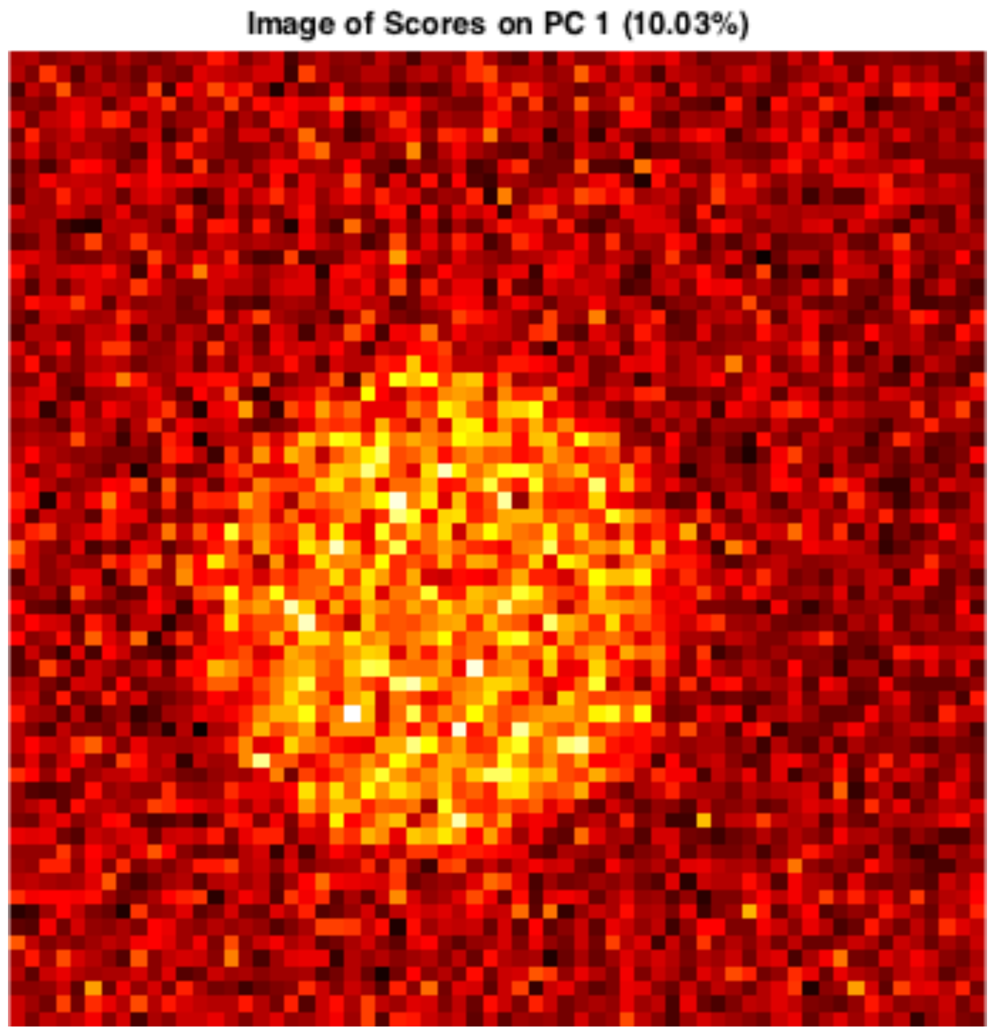


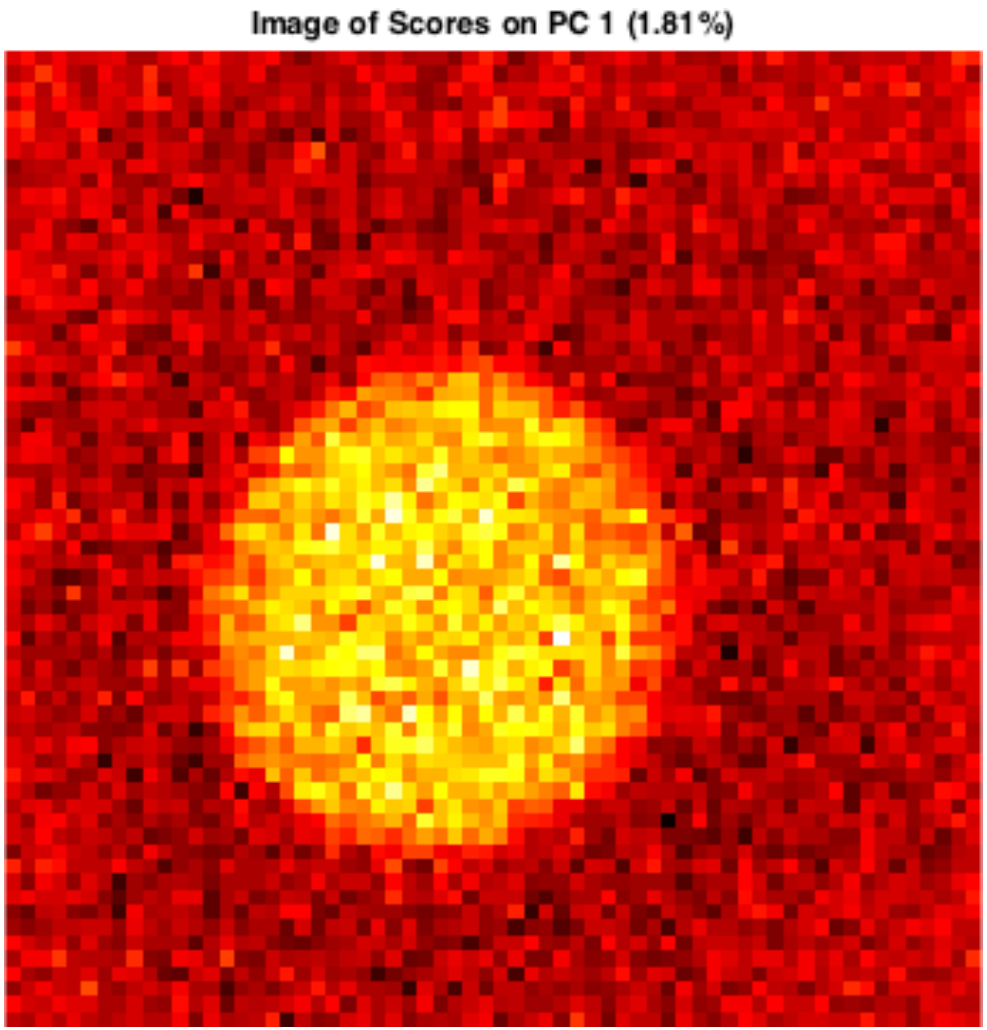
# Other Ways of Focusing on Variance of Interest

- Maximum Autocorrelation Factors (MAF) – find variance with spatial correlation
- Maximum Difference Factors (MDF) – find variance with spatial transitions (edge detection)
- Generalized Least Squares Weighting (GLS) – ignore variance from specified regions
- External Parameter Orthogonalization (EPO) – same idea as GLS

# MAF on SIMS Image of PVA




PCA



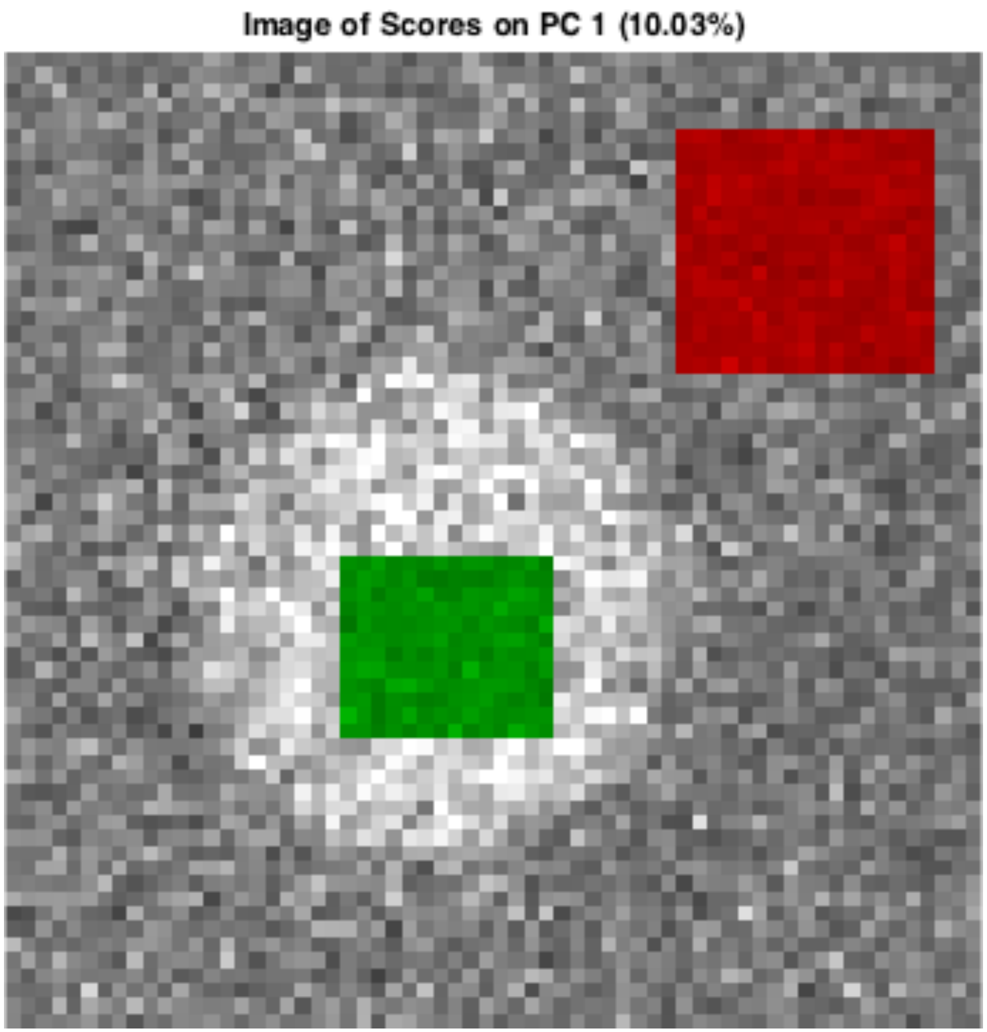

MAF

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

# Define Clutter Areas



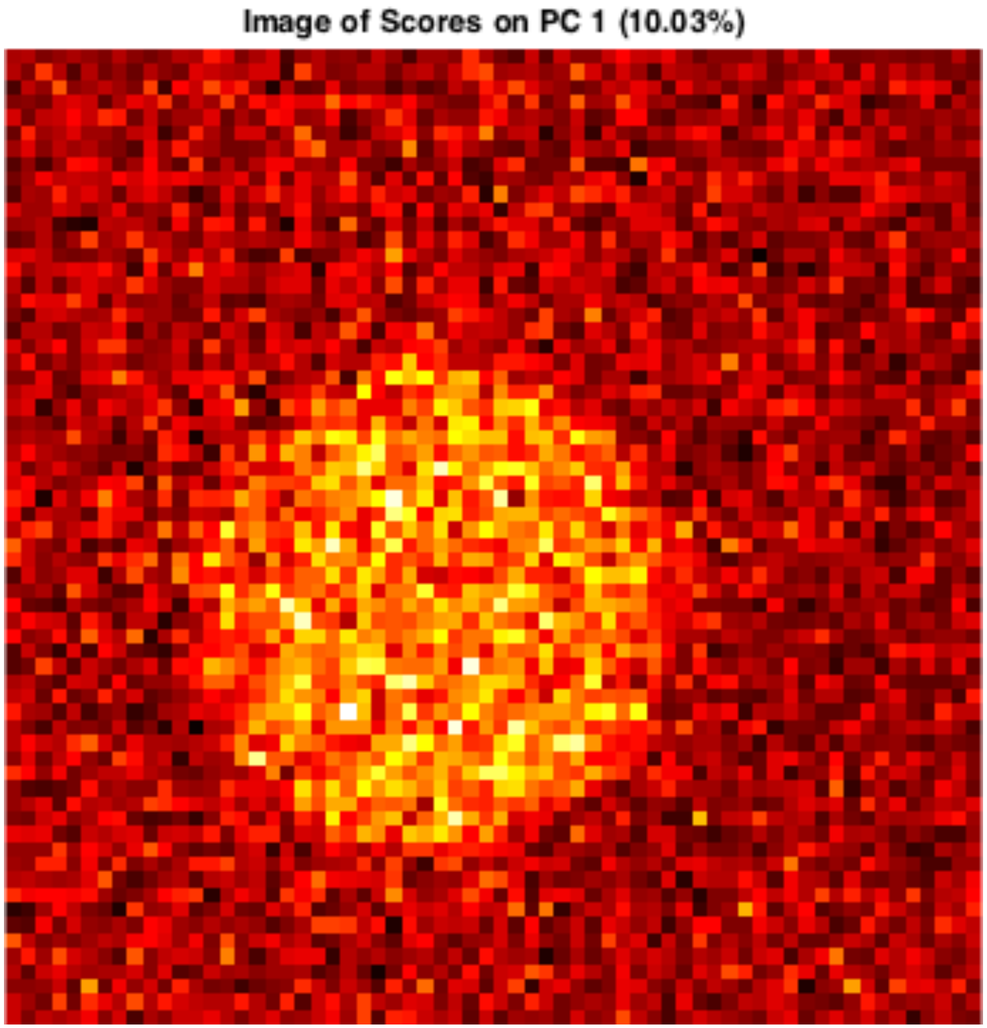

Only variation in marked areas is due to “noise”

Center each area to its own mean, then combine areas

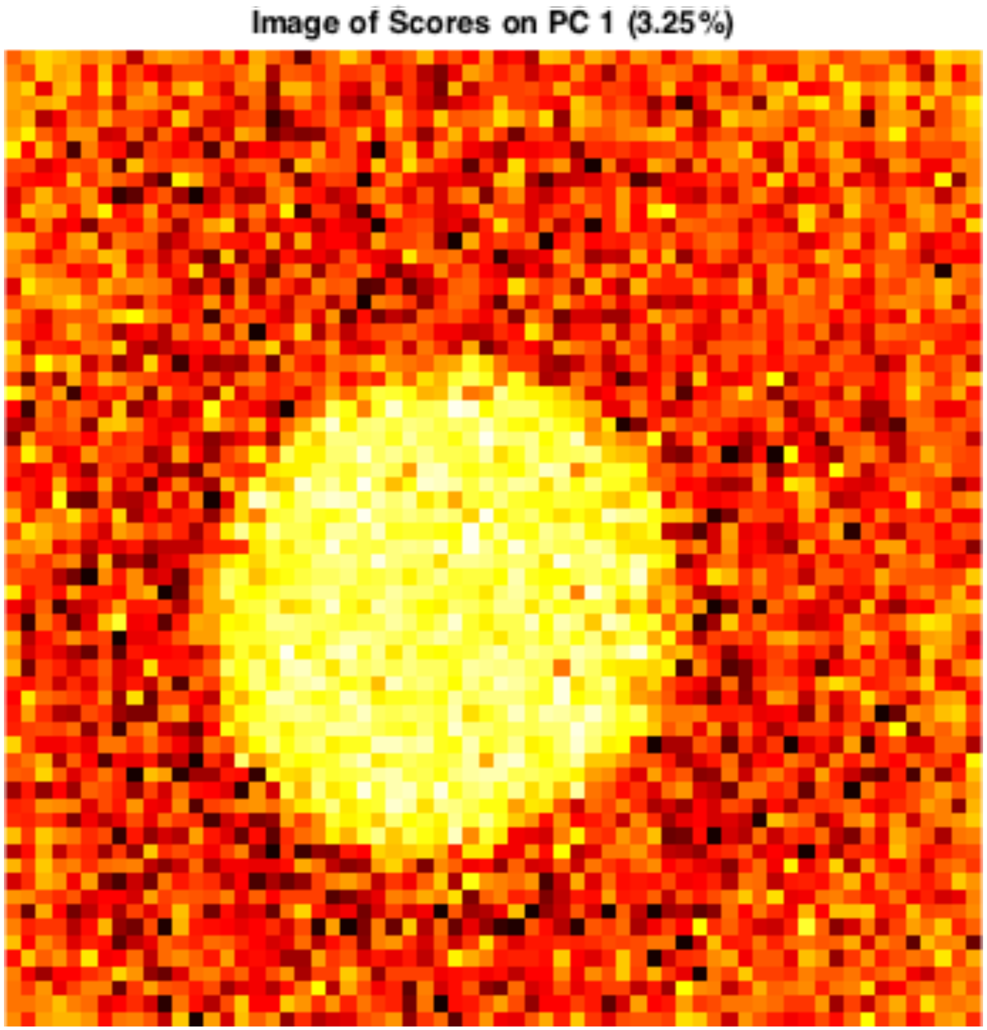
Develop GLS weighting from combined areas

EIGENVECTOR RESEARCH INCORPORATED

# GLS Filtered PVA




PCA




PCA with GLS

EIGENVECTOR RESEARCH INCORPORATED

# Cluster Analysis

- **Clustering:** Identification of natural groupings (classes) of samples without using prior knowledge of their identity – unsupervised classification
- **Agglomerative Clustering:** Start with each object as it's own cluster, then ***combine*** these into larger clusters
- **Partitional Clustering:** Start with all objects in one cluster, then ***separate*** them into smaller clusters
  - Better for image data



EIGENVECTOR RESEARCH INCORPORATED

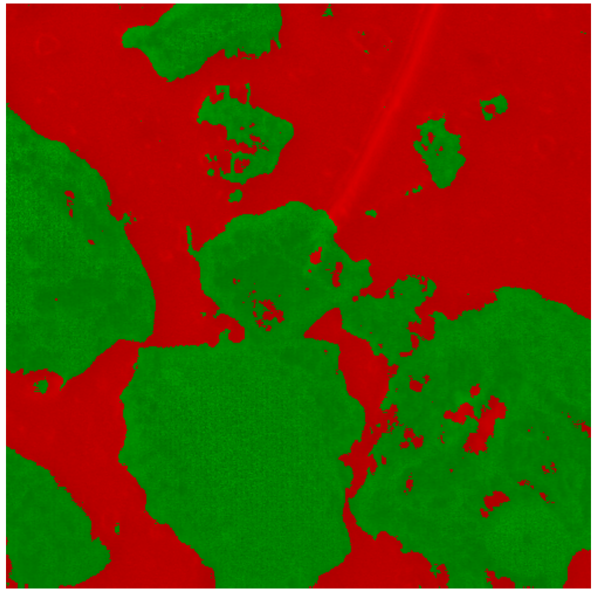
# K-Means Partitional Clustering

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

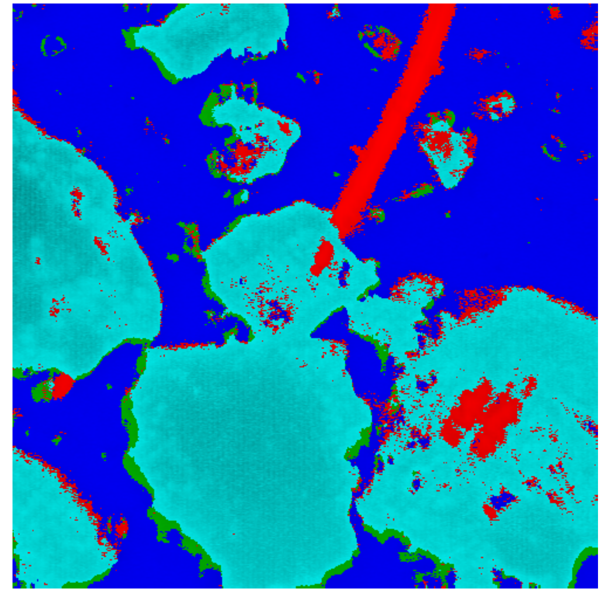

EIGENVECTOR RESEARCH INCORPORATED

# Cluster Results

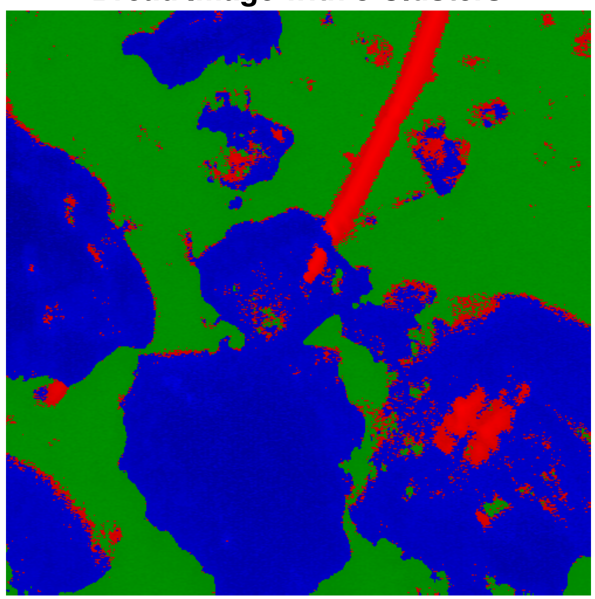
Bread Image with 2 Clusters



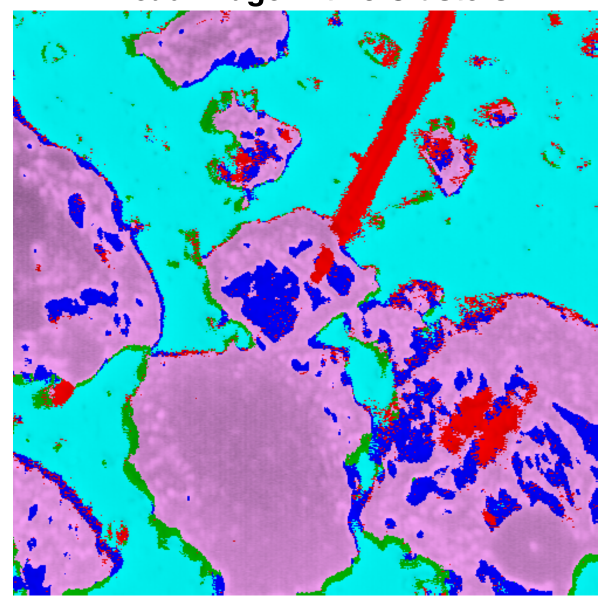
Bread Image with 4 Clusters



Bread Image with 3 Clusters



Bread Image with 5 Clusters



EIGENVECTOR RESEARCH INCORPORATED

# Conclusions

- PCA and related techniques focus on structure in the spectral (as opposed to spatial) dimension
- Condenses information from many variables, improves signal to noise
- MCR and ICA attempt to get chemically meaningful factors
- Anything that can be done with 2-way data tables can also be done with multivariate/hyperspectral images