Three-Way Factor Analysis of Large-Microscopic Hyperspectral Images: Compression and Analysis of Very Large, Small Images

Mark H. Van Benthem and Michael R. Keenan*

Sandia National Laboratories

Albuquerque, NM

*retired current address:
8346 Roney Rd.
Wolcott, NY 14590

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Overview

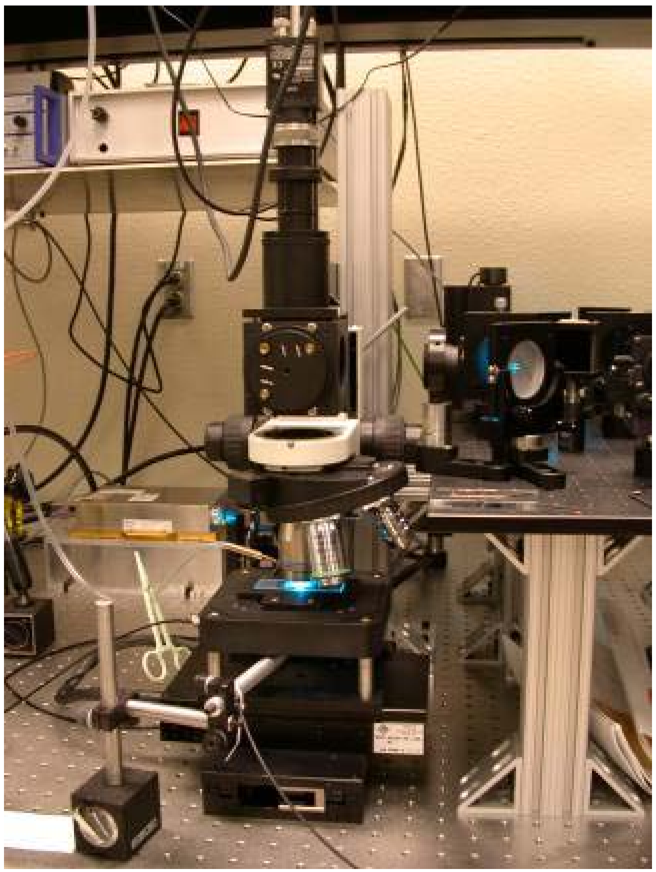
- Introduction and Motivation of Work
- Equipment and Experimental Methods
  - Hyperspectral imaging confocal microscope
  - Two- and three-way data collection
- Trilinear Analysis and PARAFAC
- Principal Components Analysis of 3-way Data
  - The Tucker1 Model and data compression
- Spatial Image Compression
  - Faster with improved noise characteristics

Motivation

- Hyperspectral microscopy is becoming commonplace in science and engineering
- Trilinear analysis of three-way data can provide rotationally unique pure components
  - Improved capability over bilinear modeling (PCA, MCR, etc.)
- Data sets are very large and can easily overwhelm researchers and users
  - Close to 1Gbyte per image
- Need improved data analysis methods to process all of the data

Hyperspectral Imaging Microscopy

3D Hyperspectral Imaging Confocal Microscope (HSI-CM)



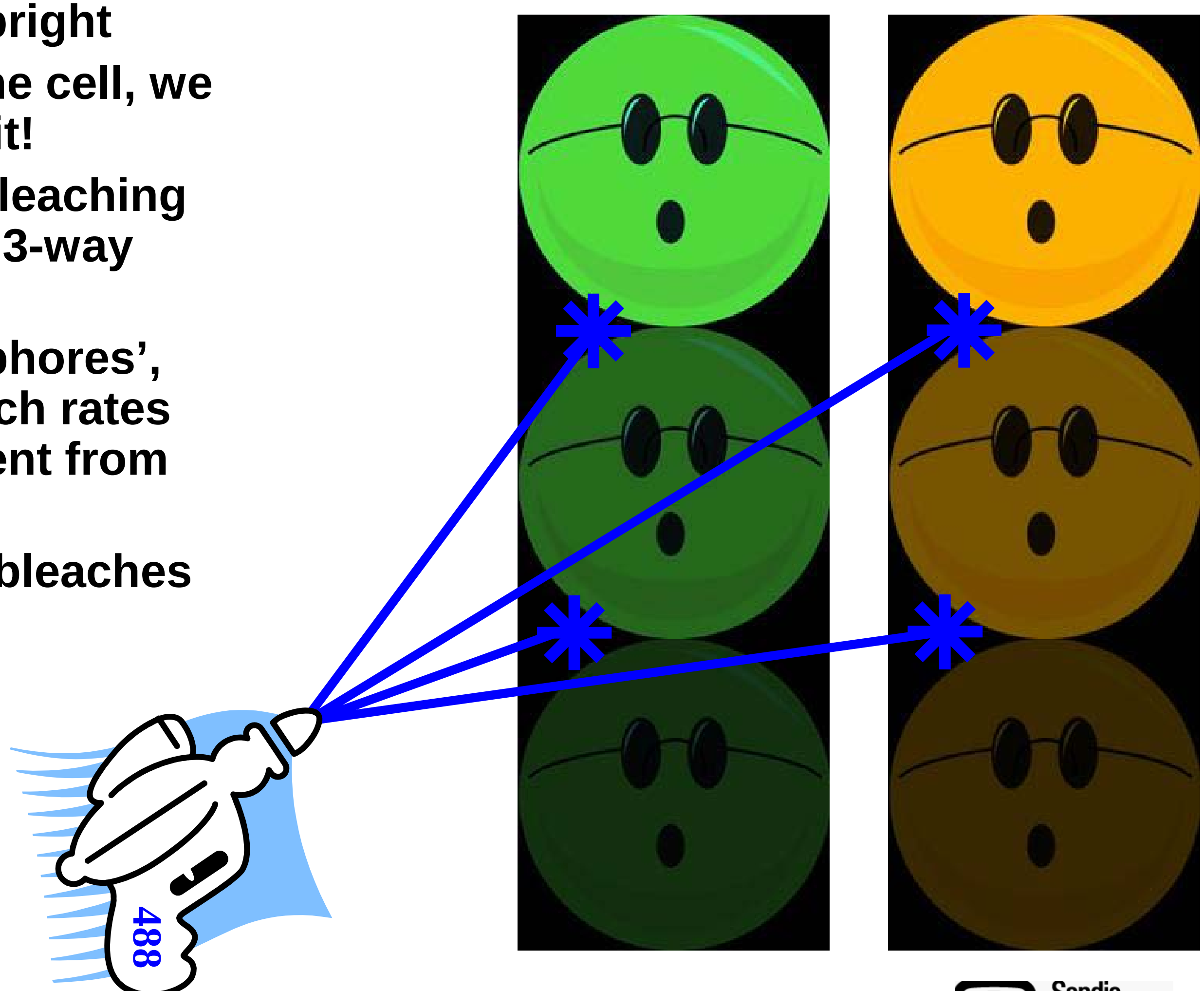
- **HSI-CM Performance Specifications:**
  - 488 nm laser excitation
  - 10x, 20x, 60x, 100x objectives
  - Lateral Resolution: 0.25 μm
  - Axial Resolution: 0.60 μm
  - Spectral range: 490-800 nm
    - 512 channels
  - Spectral resolution: 1-3 nm
  - Max acquisition rate: 8300 spectra/s
- **Collection of 2-Way Data**
  - Image-mode by spectral-mode
  - Typical Image size: 208x204 pixels
- **Collection of 3-Way Data**
  - Photobleaching

Sinclair MB, Haaland DM, Timlin JA, Jones HDT. *Appl. Opt.*, 2006; **45**: 6283-6291.

Sandia National Laboratories

Creating 3-Way Data Using Photobleaching

- Microscope laser is very bright
  - Each time we image the cell, we coincidentally bleach it!
- At the same time we are bleaching the cell, we are creating a 3-way array of data!
- Critical condition: Fluorophores’, native or introduced, bleach rates dependent must be different from one another.
- Typically collect 18 photobleaches per data set.



FRET: fluorescence resonance energy transfer.

208x204 x 512 x 18 = 391,053,312 elements

The Problems of Large Data Sets

- Multiple copies of the data representing each of the three modes may not possible
  - Common method for many PARAFAC algorithms
- Retention of Full Data Set in Available RAM
  - Even if full data can be loaded, there may be a limit on operations one can perform (e.g., transposition)
- Time required to perform analysis is extensive
- Potential solutions:
  - Tucker1 and Tucker3 models of data
    - Essentially, PCA in three modes
  - Image Compression

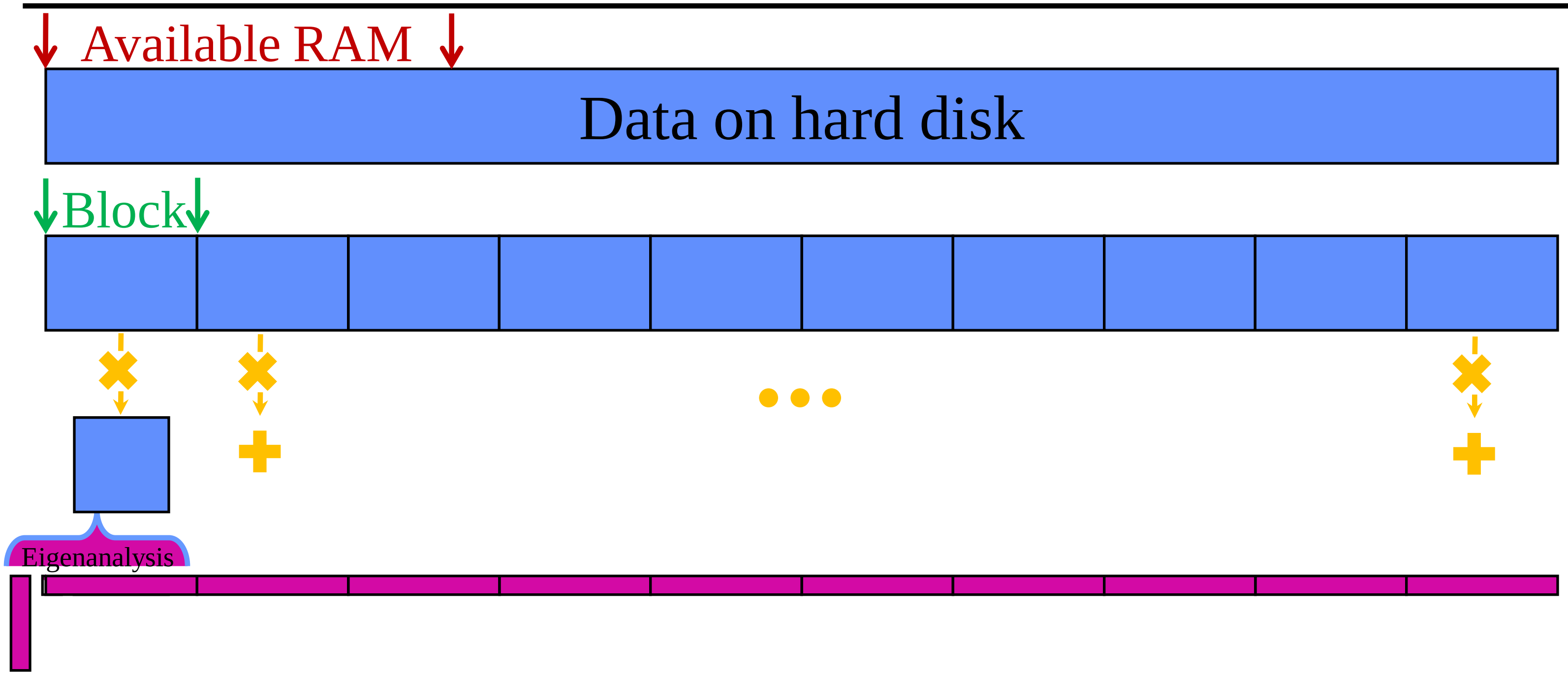
Sandia National Laboratories

Principal Component Analysis (PCA)

- Data reduction technique
  - Reduces the dimensionality of or “factors” the data to the appropriate number of components
  - Creates an ordered set of orthogonal “scores,” **T**, and “loadings,” **P**.
    - These are vectors representing the image and spectral modes ordered by decreasing variance contribution to the data set.
- Related to eigenanalysis, which is the usual method of computation
  - For a large data set, compute the small cross-product matrix $\mathbf{Z} = \mathbf{F}^{T}\mathbf{F}$ and then solves the eigenvalue problem $\mathbf{ZP} = \mathbf{PD}$, where **D** is the diagonal matrix of eigenvalues
  - Then, sort the eigenvalues by size, order the vectors in **P** accordingly, select the appropriate number of vectors
  - Next solve for **T** using $\mathbf{T} = \mathbf{FP}$
  - And the result is the model $\mathbf{F} = \mathbf{TP}^{T} + \mathbf{E}$

Sandia National Laboratories

"Out-of-Core" Block PCA for Large Data



- Problem: Data too large to fit into main memory
- Solution:
  - Identify blocks of data for sequential access
  - Load blocks & form cross-product on small side, in-turn
  - Perform eigenanalysis on summed cross-products to obtain scores
  - Reload blocks and project into loading space to obtain loadings

Three-Way Data and Trilinear Methods

- Three-way fluorescence data that follow a trilinear model
  - The fluorescence intensity is collected as a function of three independent parameters
  - Intensity varies linearly as a function of each parameter
- The trilinear model:
  - $\mathbf{F} = \otimes(\mathbf{S},\mathbf{T},\mathbf{C}) + \mathbf{E}$, which is a three-way array
- Provides rotationally unique decompositions
  - Given appropriate data rank structure

$\otimes$ triple-product operator

Sandia National Laboratories

Cyclic Permutation of 3-Way Data

- **Given data in $\mathbb{F}$ is "matricized" as F, it has spectral-mode in rows and bleach (outer loop)-image (inner loop) modes in columns**
- **In Matlab this is accomplished using *transpose* and *reshape* (index reassignment)**

$\mathbb{F} \Rightarrow$ reshape(F, $N_{spectral}$, N_{bleach}*N_{image})

F = reshape(transpose(F) ,N_{image}, $N_{spectral}$*N_{bleach})

F = reshape(transpose(F), N_{bleach}, N_{image}*$N_{spectral}$)

F = reshape(transpose(F), $N_{spectral}$, N_{bleach}*N_{image})

Costs: *transpose* \$\$\$, *reshape* ¢¢¢

Ref. Jiang, J. H.; Wu, H. L.; Li, Y.; Yu, R. Q. *J. Chemom.* ***1999, 13, 557.***

Three-Way PCA

- Three-Mode PCA
  - **F** is the three-way data array arranged as a matrix
  - Three sets of orthogonal "loadings" **V**, **U**, and **W**
  - Model the spectral, image and photobleach modes
  - Core array **G** which models the variance of the data in array **F**
- The model: $\mathbf{F} = \mathbf{V}\mathbf{G}(\mathbf{W}\otimes\mathbf{U})^{T}$
- Tucker3: Solve the equation in a "least squares" sense using an alternating least squares routine
- Tucker1: Solve the equation for each mode independently

$\otimes$ is the Kronecker product.

Tucker3 Model

- **Minimize:** $\left\| \mathbf{F} - \mathbf{V}\mathbf{G}\left(\mathbf{W}\otimes\mathbf{U}\right)^{T} \right\|_2^2$ *s.t.* **(V,U,W)** orthonormal
- **Perform eigenanalysis on array by iteratively solving**
  - **Spectral mode: Let $\mathbf{Z}=\mathbf{F}(\mathbf{W}\otimes\mathbf{U})$, solve $(\mathbf{Z}\mathbf{Z}^{T})\mathbf{V} = \mathbf{V}\mathbf{D}$**
  - **Image mode: Let $\mathbf{Z}=\mathbf{F}(\mathbf{V}\otimes\mathbf{W})$, solve $(\mathbf{Z}\mathbf{Z}^{T})\mathbf{U} = \mathbf{U}\mathbf{D}'$**
  - **Bleach mode: Let $\mathbf{Z}=\mathbf{F}(\mathbf{U}\otimes\mathbf{V})$, solve $(\mathbf{Z}\mathbf{Z}^{T})\mathbf{W} = \mathbf{W}\mathbf{D}''$**
- **Run until convergence**
- **Solve for Core array: $\mathbf{G} = \mathbf{V}^{T}\mathbf{F}(\mathbf{W}\otimes\mathbf{U})$**
- **Computationally and time intensive**
- **Ever more challenging if you can't load the full data set into memory.**

Ref. Andersson CA and Bro R, *Chemometrics Intell. Lab. Syst., 1998;* ***42: 93-103.***

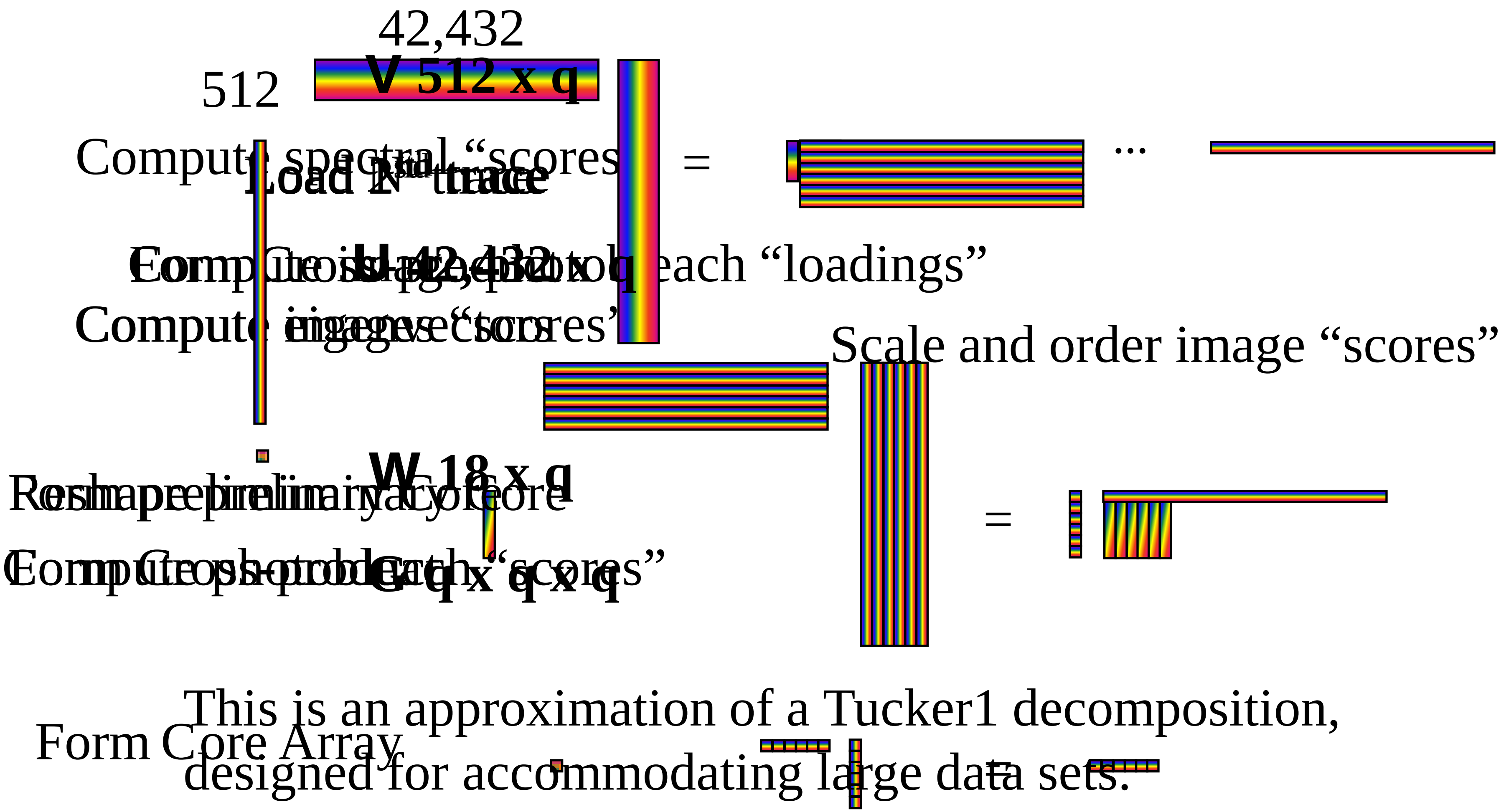
Tucker1 Model

- Perform eigenanalysis on each mode in array
  - Spectral mode: $(\mathbf{F}\mathbf{F}^{\mathsf{T}})\mathbf{V} = \mathbf{V}\mathbf{D}$
  - Image mode: $(\mathbf{F}\mathbf{F}^{\mathsf{T}})\mathbf{U} = \mathbf{U}\mathbf{D}'$
  - Bleach mode: $(\mathbf{F}\mathbf{F}^{\mathsf{T}})\mathbf{W} = \mathbf{W}\mathbf{D}''$
- Solve for Core array:
  - $\mathbf{G} = \mathbf{V}^{\mathsf{T}}\mathbf{F}(\mathbf{W}\otimes\mathbf{U})$
- Again, what happens if you can't load the full data set into memory?
  - Or if the cross-product is too large, as in our image mode? Size of (208x204) → 42,432 x 42,432!

Fast Tucker1 Method

- **Perform eigenanalysis on one manageable mode, choosing an appropriate number of factors**
  - **Spectral mode:** $(\mathbf{F}\mathbf{F}^{T})\mathbf{V} = \mathbf{V}\mathbf{D}$
  - **If necessary load one photobleach layer at a time**
- **Compute the scores matrix (array) H on the remaining modes**
  - $\mathbf{H} = \mathbf{V}^{T}\mathbf{F}$
- **Permute scores matrix in the same manner as the data array**
  - **(H→H as F→F, and H→H as F→F)**
- **Perform eigenanalysis on the remaining modes in the scores array analogous to Tucker1 on full data**

Fast Tucker1 Method Schema



This is an approximation of a Tucker1 decomposition, designed for accommodating large data sets.

Van Benthem MH, Keenan MR, J. Chemometrics, 2008; 22: 345-354.

PARAFAC-ALS Using Core Matrix Only

Step 1: $\hat{\tilde{\mathbf{S}}} = \mathbf{G}\left(\tilde{\mathbf{T}} \odot \tilde{\mathbf{C}}\right)\left(\tilde{\Omega}\right)^{-1}$

$$\tilde{\Omega} = \left(\tilde{\mathbf{T}} \odot \tilde{\mathbf{C}}\right)^{\mathrm{T}}\left(\tilde{\mathbf{T}} \odot \tilde{\mathbf{C}}\right) = \left(\tilde{\mathbf{C}}^{\mathrm{T}}\tilde{\mathbf{C}}\right) * \left(\tilde{\mathbf{T}}^{\mathrm{T}}\tilde{\mathbf{T}}\right)$$

Step 2: $\hat{\tilde{\mathbf{C}}} = \mathbf{G}\left(\hat{\tilde{\mathbf{S}}} \odot \tilde{\mathbf{T}}\right)\left(\tilde{\Xi}\right)^{-1}$

$$\tilde{\Xi} = \left(\tilde{\mathbf{T}}^{\mathrm{T}}\tilde{\mathbf{T}}\right) * \left(\hat{\tilde{\mathbf{S}}}^{\mathrm{T}}\hat{\tilde{\mathbf{S}}}\right)$$

Step 3: $\hat{\tilde{\mathbf{T}}} = \mathbf{G}\left(\hat{\tilde{\mathbf{C}}} \odot \hat{\tilde{\mathbf{S}}}\right)\left(\tilde{\Pi}\right)^{-1}$

$$\tilde{\Pi} = \left(\hat{\tilde{\mathbf{S}}}^{\mathrm{T}}\hat{\tilde{\mathbf{S}}}\right) * \left(\hat{\tilde{\mathbf{C}}}^{\mathrm{T}}\hat{\tilde{\mathbf{C}}}\right)$$

After convergence:

$$\hat{\mathbf{S}} = \mathbf{V}\hat{\tilde{\mathbf{S}}} \qquad \hat{\mathbf{C}} = \mathbf{U}\hat{\tilde{\mathbf{C}}} \qquad \hat{\mathbf{T}} = \mathbf{W}\hat{\tilde{\mathbf{T}}}$$

Sandia National Laboratories

PARAFAC-ALS Using 3-Mode PCA Factors

$$\text{Step 1: } \hat{\mathbf{S}} = \left(\mathbf{V}\left(\mathbf{G}\left(\mathbf{W}^{\mathrm{T}}\mathbf{T} \odot \mathbf{U}^{\mathrm{T}}\mathbf{C}\right)\right)\right)(\boldsymbol{\Omega})^{-1}$$

$$\boldsymbol{\Omega} = (\mathbf{T} \odot \mathbf{C})^{\mathrm{T}}(\mathbf{T} \odot \mathbf{C}) = \left(\mathbf{C}^{\mathrm{T}}\mathbf{C}\right) * \left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right)$$

$$\text{Step 2: } \hat{\mathbf{C}} = \left(\mathbf{U}\left(\mathbf{G}\left(\mathbf{V}^{\mathrm{T}}\hat{\mathbf{S}} \odot \mathbf{W}^{\mathrm{T}}\mathbf{T}\right)\right)\right)(\boldsymbol{\Xi})^{-1}$$

$$\boldsymbol{\Xi} = \left(\mathbf{T}^{\mathrm{T}}\mathbf{T}\right) * \left(\hat{\mathbf{S}}^{\mathrm{T}}\hat{\mathbf{S}}\right)$$

$$\text{Step 3: } \hat{\mathbf{T}} = \left(\mathbf{W}\left(\mathbf{G}\left(\mathbf{U}^{\mathrm{T}}\hat{\mathbf{C}} \odot \mathbf{V}^{\mathrm{T}}\hat{\mathbf{S}}\right)\right)\right)(\boldsymbol{\Pi})^{-1}$$

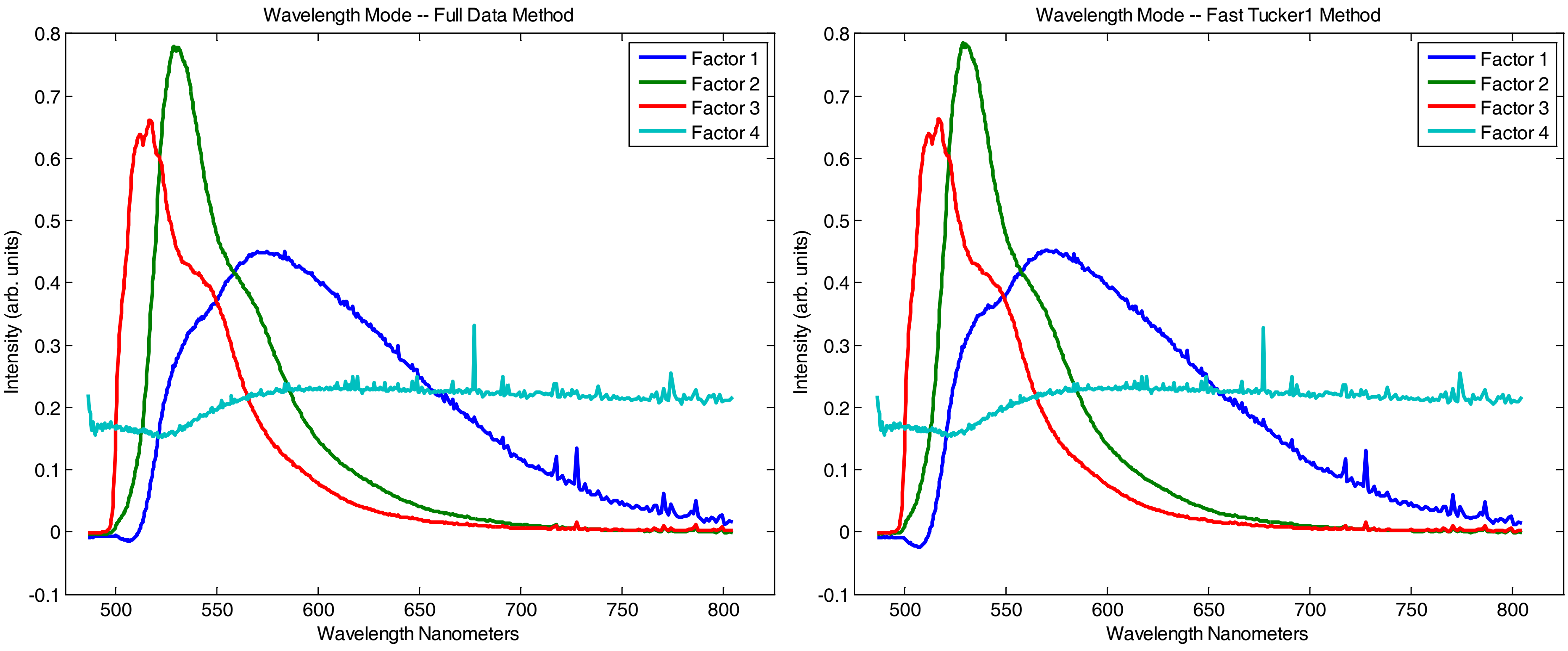
$$\tilde{\boldsymbol{\Pi}} = \left(\hat{\mathbf{S}}^{\mathrm{T}}\hat{\mathbf{S}}\right) * \left(\hat{\mathbf{C}}^{\mathrm{T}}\hat{\mathbf{C}}\right)$$

Sandia National Laboratories

Performance Comparison on HSI-CM Data

- HEK 293 cells
  - Transiently transfected with GFP and YFP
- HSI-CM data collection
  - Original size 199×204 image pixels by 512 wavelength channels by 18 photobleach steps
    - Image reduced to 159×204 due to some saturation
- Pretreatment
  - Remove cosmic spikes
  - Subtracting simple offset added by the EMCCD
  - Poisson scaling prior to 3-way PCA
- Performed PARAFAC on full data set and compressed with Fast Tucker1 algorithm
  - Matlab Version7.3.0.267 (R2006b)
  - Red Hat Linux, 64 bit Enterprise 4 version
  - Dell, dual core, 3.2 GHz Xeon processor w/ 6.0 GB RAM

Spectral Mode Results



Computation Times

4702 seconds -- 79 seconds

Sandia National Laboratories

Photobleach Mode Results

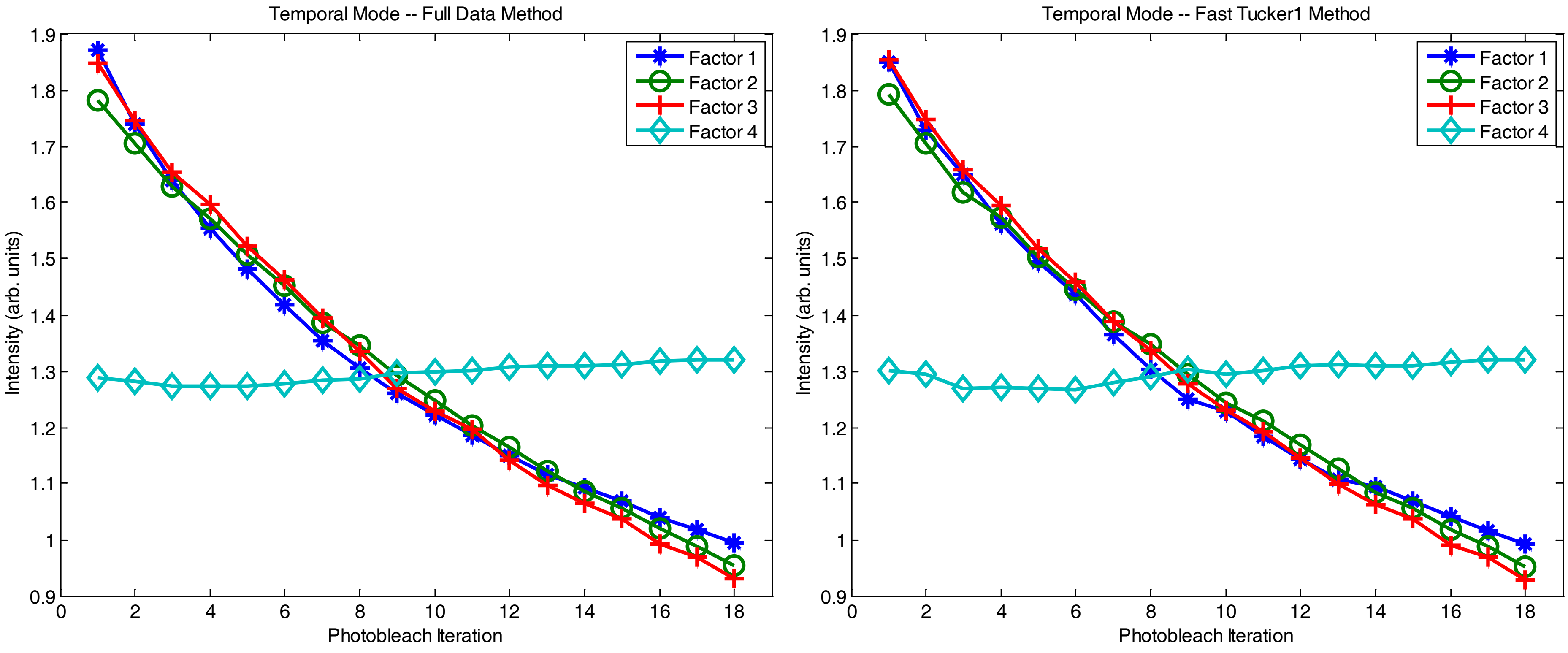
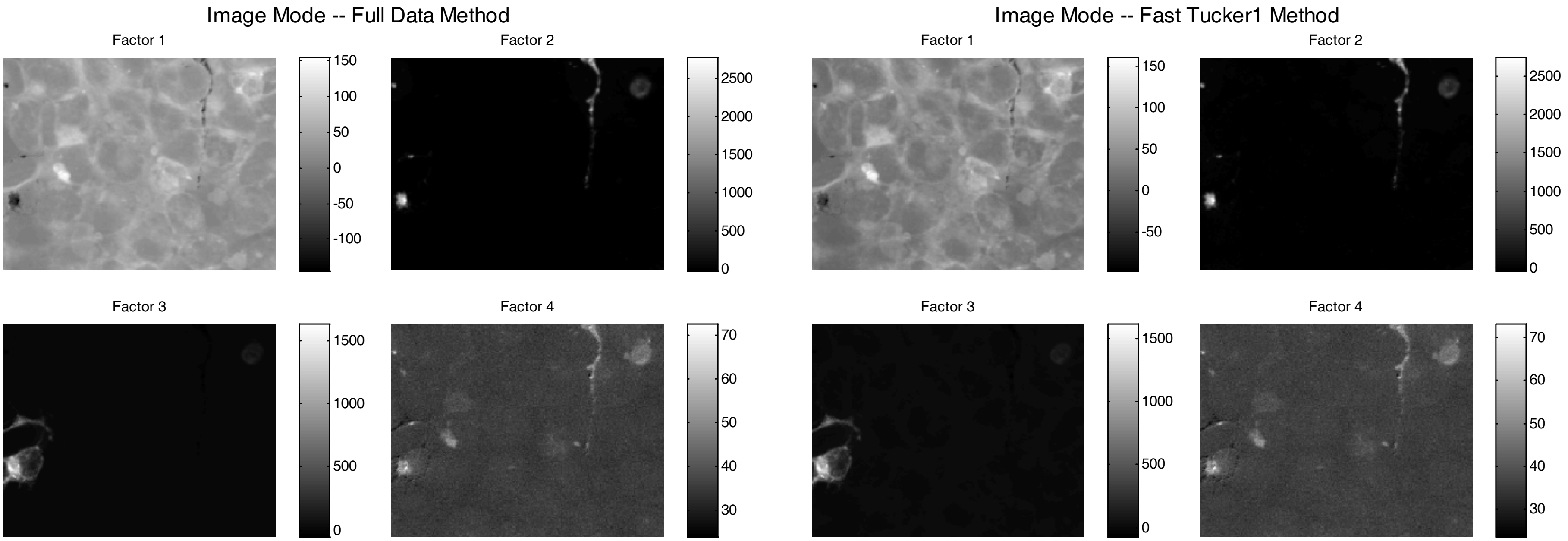


Image Mode Results

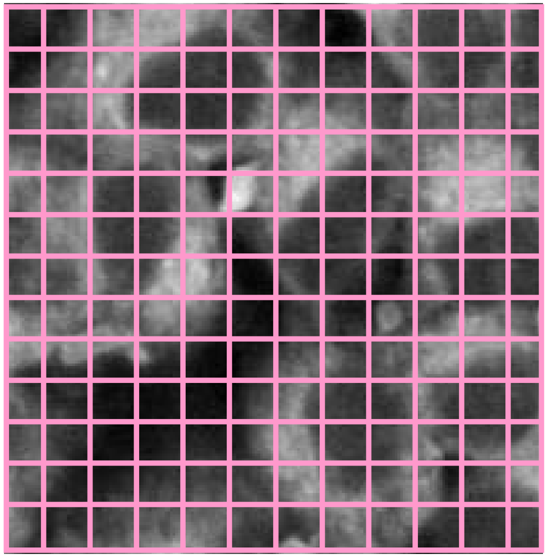


Data Compression

- Can be simple and easy to apply (e.g., binning)
- Can achieve good analysis results with very high compression factors
- Can improve signal-to-noise characteristics of data
  - Random noise adds destructively
  - Signal adds constructively, enhancing signal
  - Binning preserves Poisson noise (Poisson adds as Poisson) and read noise (variances add)
- Subsequent analysis proceeds much faster!

Spatial Binning Compression of a Cell Image

Sum all of the pixels
in each square



Original Image
208×204 pixels

Compression factors of
16 and 17 (total 272)
chosen for convenience

Compressed Image
13×12 pixels

Compression doesn’t have to preserve the image “quality.”
But, if you stand way back…

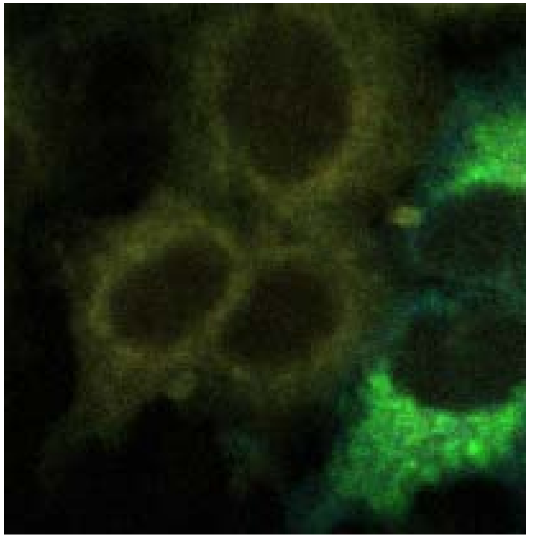
Sandia
National
Laboratories

Three-Way Cell Data

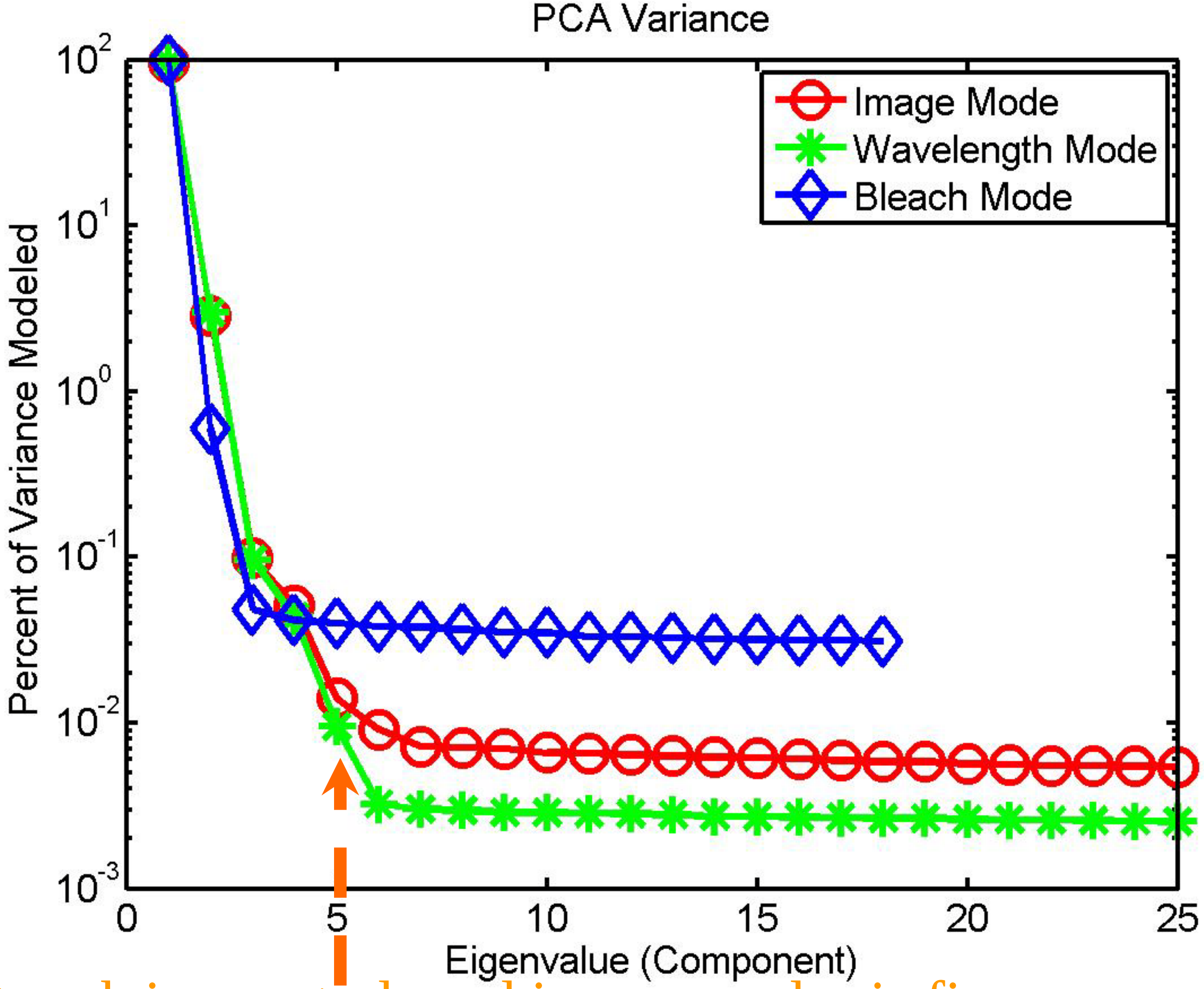
- Collected 18 consecutive photobleach images of same cells previously examined
- Preprocessing
  - Remove cosmic ray-induced "spikes"
  - Remove dark shape—rank one PCA component
    - Wavelength dependent spectral response of EMCCD
  - Trimmed first six Spectral-mode data points
  - Subtract offset for each image pixel
    - Baseline correct using known zero-signal elements
  - Binning compression in image-mode as previously described
  - Poisson and read-noise scale in spectral-mode
- Analyzed with a PARAFAC-ALS routine written in-house

Hyperspectral Cell Image

- RGB image
  - Human A549 pulmonary type II epithelial cell
  - Labeled with GFP and YFP
  - Collected with HSI-CM
- Produced by integrating over 3 wavelength bands
  - Chosen for responses of GFP, YFP and cell autofluorescence.

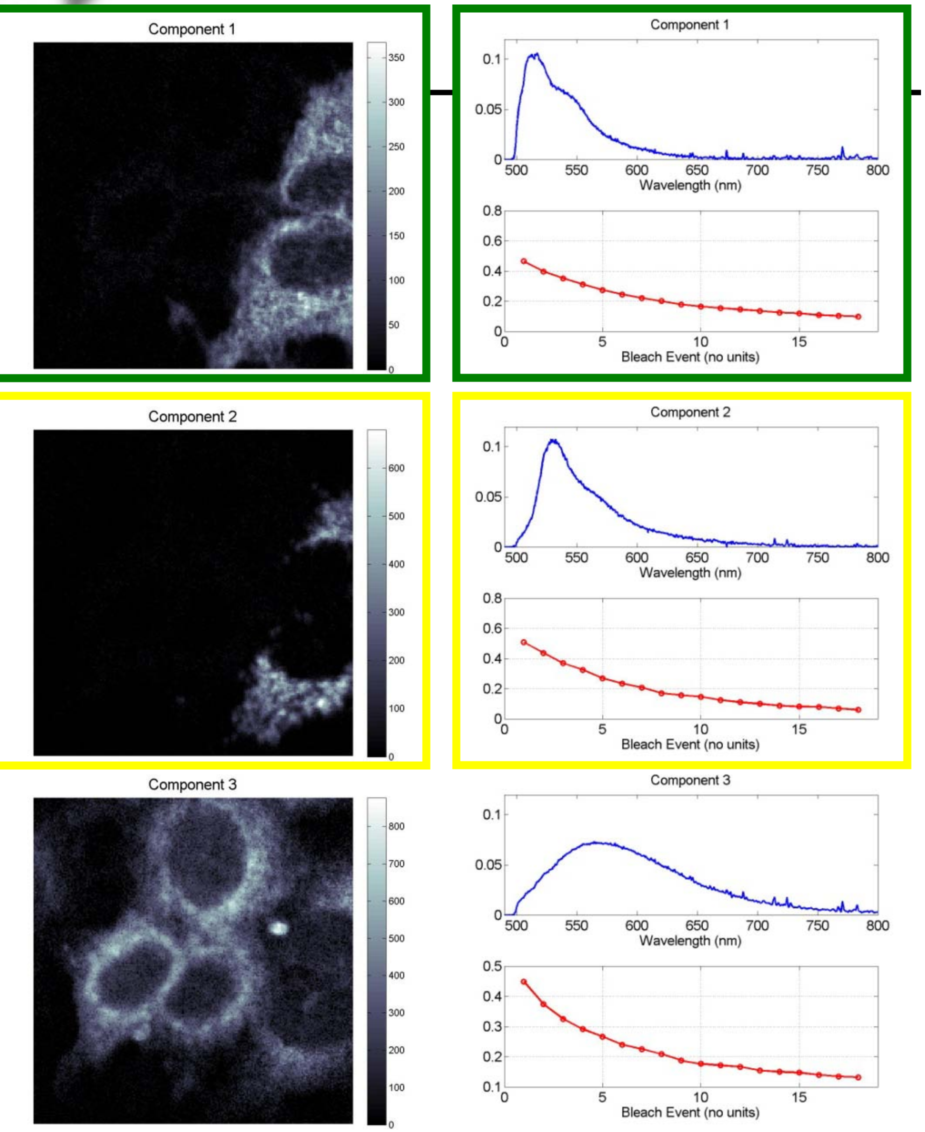


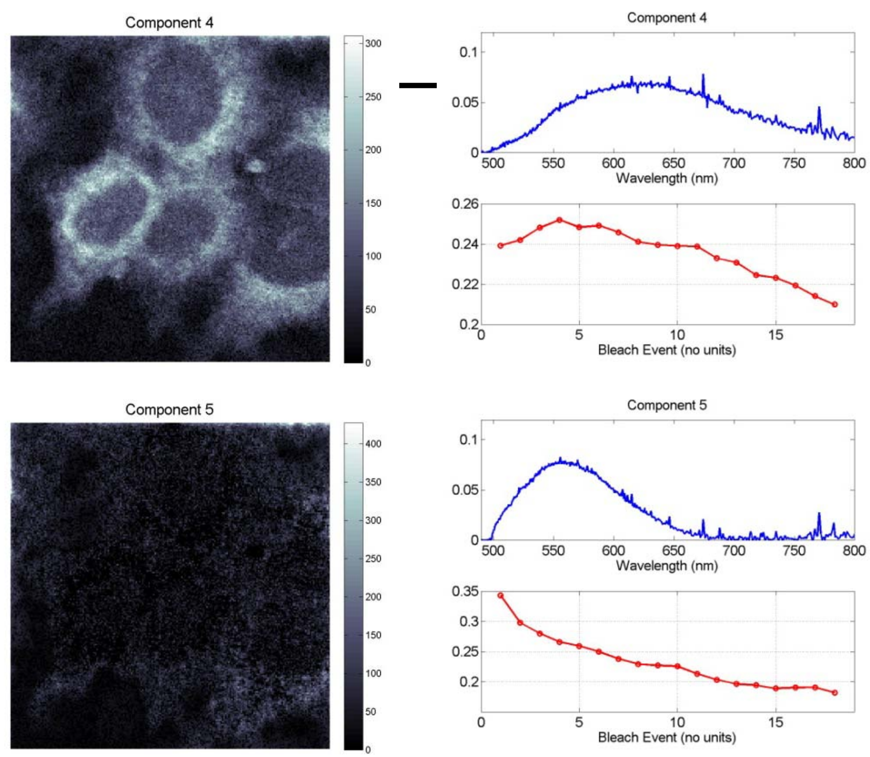
HSI Cells Three-Mode Rank Estimate



Largest rank in spectral- and image-modes is five.

GFP-YFP-AF PARAFAC Model





PARAFAC easily finds GFP and YFP fluors and returns their photobleach curves and their cell locations.

Van Benthem, M. H.; Keenan, M. R.; Davis, R.; Liu, P.; Jones, H. D. T.; Haaland, D. M.; Sinclair, M. B.; Brasier, A. R. *Journal of Chemometrics* ***2008 22, 491.***

Sandia National Laboratories

Recap and Summary

- Presented methods of trilinear analysis for multivariate analysis of large hyperspectral images
  - Least squares for linear models
- Fast Tucker1 method is very fast way to compress large data sets
  - Approximation to the standard Tucker1 model
  - Easy to implement follow-on analyses after compression
- Compression of data can take many forms
  - Binning large hyperspectral images is simple and works well for linear additive data
- PARAFAC is a powerful tool for three-way analysis
  - Data must follow the trilinear model

Acknowledgements

- Sandia National Laboratories
  - Ryan Davis, David Haaland, Howland Jones, Jeri Timlin, Mike Sinclair, and Chris Stork
- University of Texas Medical Branch
  - Allan Brasier, Ping Liu