SAND2018-6088C

PARAFAC2: A Core Perspective

TriCAP 2018

June 15, 2018, 0900

Mark H. Van Benthem

Sandia National Laboratories, Albuquerque, NM

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Overview

- Introduction and motivation for research
- Relevant prior work
- PARAFAC2 and Core-PARAFAC2 Algorithms
- Imposing Nonnegativity
- Analysis on Simulated Data
- Analysis on Fluorescence Data
- Results
- Conclusions and future directions

Introduction

- Numerous R&D areas employ fluorescence lifetime to characterize fundamental molecular properties.
- Multiway fluorescence measurements provide greater analytical power than single modalities
- Two-way wavelength-time matrices (WTM) collected at different excitation wavelengths can produce three-way data
  - Time-Resolved Excitation Emission Matrix (TREEM).
- PARAFAC is an excellent data analytical tool for fluorescence data, but is limited when the data do not meet the strict expectations of trilinearity
- PARAFAC2 can accommodate shifts or scale differences in one modality and overcome restrictions
- Core PARAFAC2 allows easy implementation of nonnegativity constraints in all three modes

Motivation

- PARAFAC2 is an important multiway analysis tool when PARAFAC is inappropriate.
  - One mode has axes that are not aligned or inconsistent
  - Multiple hyphenated chromatographic profiles are a prime example
- PARAFAC2 can be slow, and difficult to implement with large data sets
- Nonnegativity constraints are difficult to apply in all three data modes
  - "…, it is impossible to impose non-negativity constraints on the [varying mode],…"*
- Improving PARAFAC2 may facilitate its broader uses in data analyses

*J. M. Amigo, et al., "Solving GC-MS problems with PARAFAC2," TrAC, Trends Anal. Chem. 27(8), 714-725 (2008)

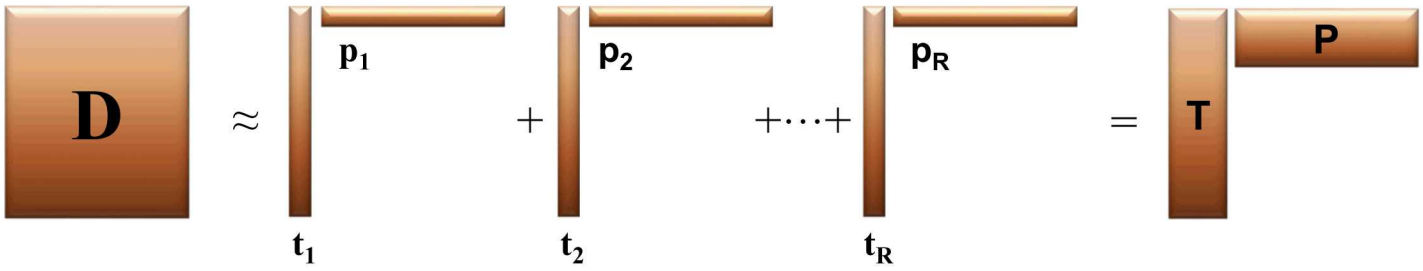
Relavent Published Research

- H. A. L. Kiers, J. M. F. TenBerge and R. Bro, "PARAFAC2 - Part I. A direct fitting algorithm for the PARAFAC2 model," J. Chemom. 13(3-4), 275-294 (1999)
- R. Bro, C. A. Andersson and H. A. L. Kiers, "PARAFAC2 - Part II. Modeling chromatographic data with retention time shifts," J. Chemom. 13(3-4), 295-309 (1999)
- J. M. Amigo, et al., "Solving GC-MS problems with PARAFAC2," TrAC, Trends Anal. Chem. 27(8), 714-725 (2008)
- M. H. Van Benthem and M. R. Keenan, "Tucker1 model algorithms for fast solutions to large PARAFAC problems," J. Chemom. 22(5), 345-354 (2008)

Recent Publications Describing PARAFAC2 with Constraints

- A. Afshar, et al., "COPA: Constrained PARAFAC2 for Sparse & Large Datasets," in eprint arXiv:1803.04572, ARXIV (2018).
- J. E. Cohen and R. Bro, "Nonnegative PARAFAC2: a flexible coupling approach," in eprint arXiv:1802.05035, ARXIV (2018).

Two-Way Analysis Methods

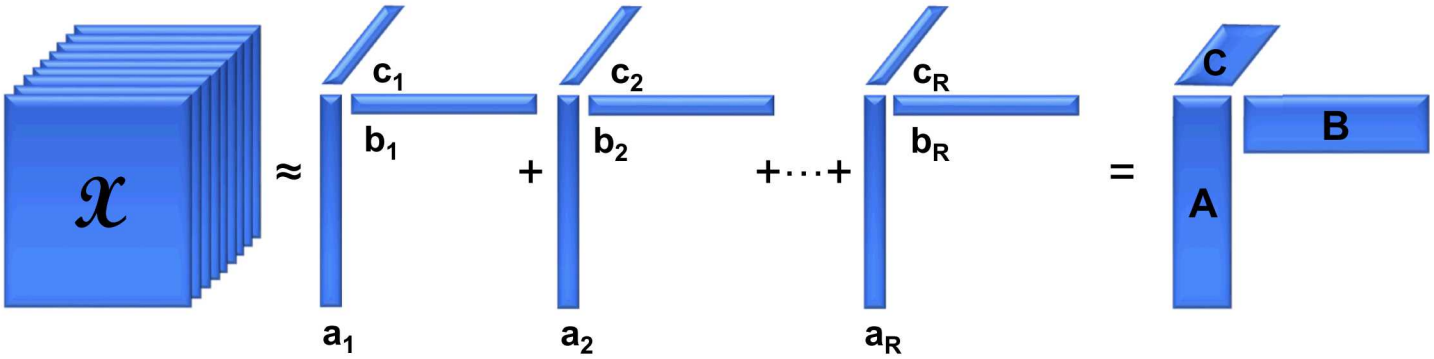


- Principal Component Analysis (PCA)
  - Given a matrix containing data, D, as a first step in many analyses we want principal components

    $$\mathbf{D} \cong \mathbf{TP}^{\mathbf{T}}$$

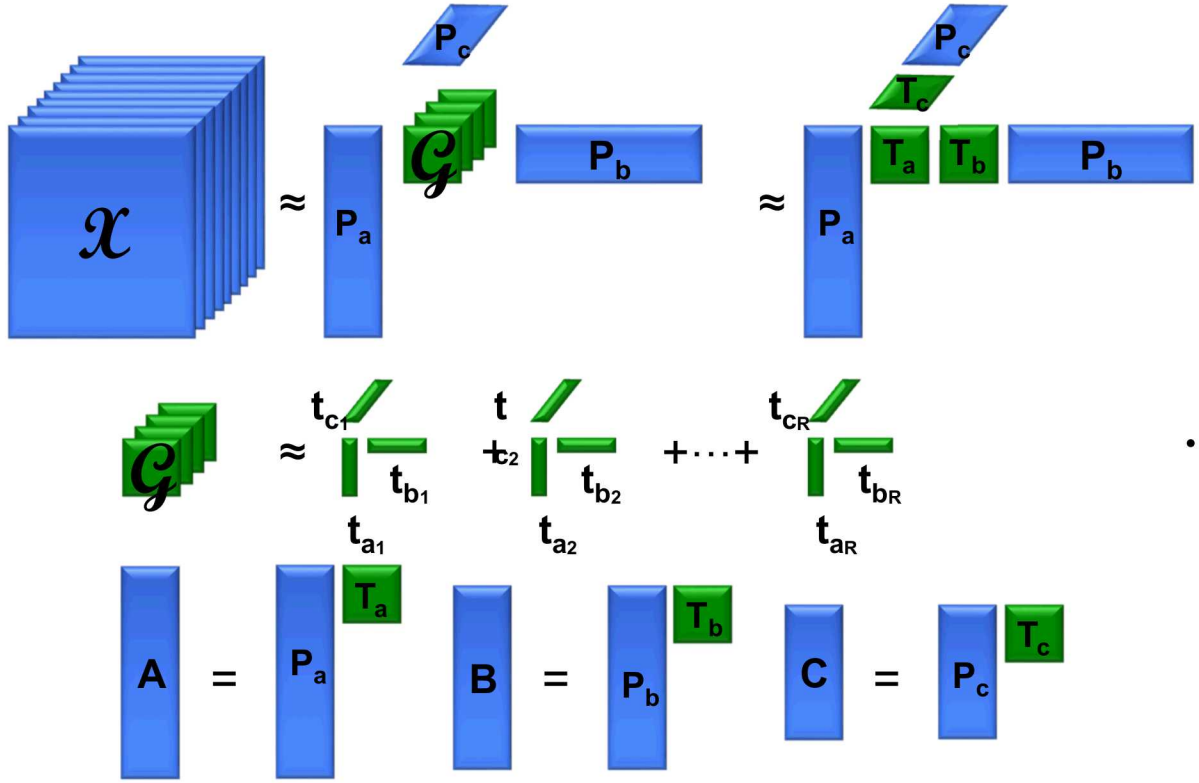
  - Such that T and P are an orthogonal basis sets, that is a reduced dimensional representation of D, with ordered maximized variance.
    - T is orthogonal (scores); P is orthonormal (loadings).
- Multivariate Curve Resolution (MCR)
  - Impose constraints on solution space

Tensor Factorization-PARAFAC



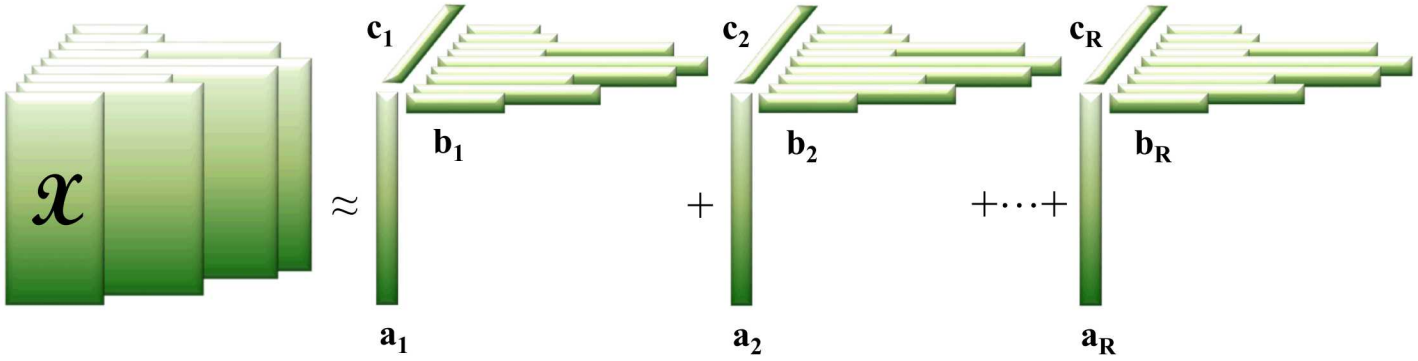
- Tensor factorizations of multi-way data
  - Parallel factor analysis (PARAFAC)
  - Nonnegative tensor factorization (NTF)
  - Similar in idea to least squares matrix techniques: principal component analysis (PCA), singular value decomposition (SVD), multivariate curve resolution (MCR)
- When applied to a data array, data are modeled as a mixture of factors, each with its own triad of signature factors

Tensor Factorization-Core PARAFAC



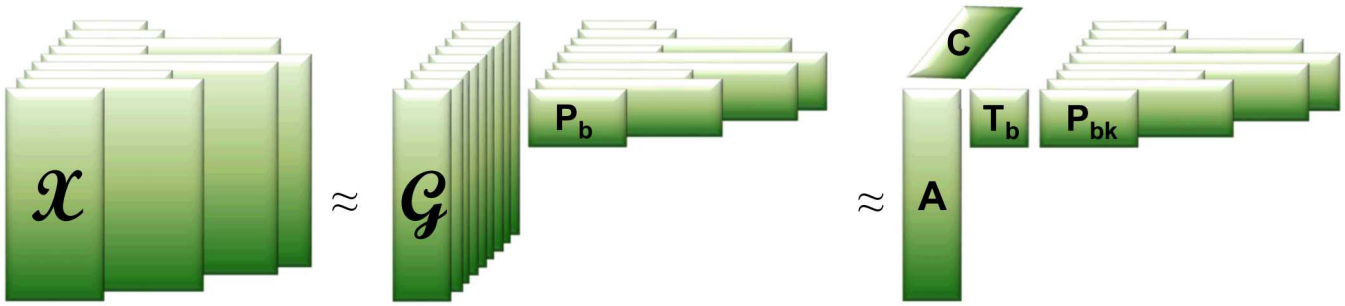
- Tensor factorizations of multi-way data
  - Tensor is initially factored using a 3-way orthogonal decomposition to generate orthogonal factors and core
  - Parallel factor analysis (PARAFAC) is performed on core to generate rotation-matrix factors

Tensor Factorization-PARAFAC2

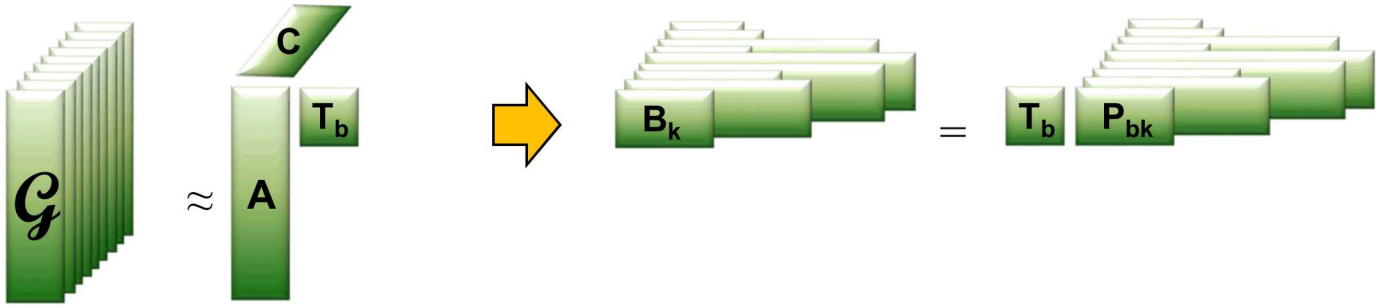


- PARAFAC2 permits factorization of unaligned or different length signals
- This data appears to not follow a model suitable for PARAFAC
  - Temporal alignment of sinus rhythms not trivial
  - Lack of alignment may prevent use of standard PARAFAC algorithm

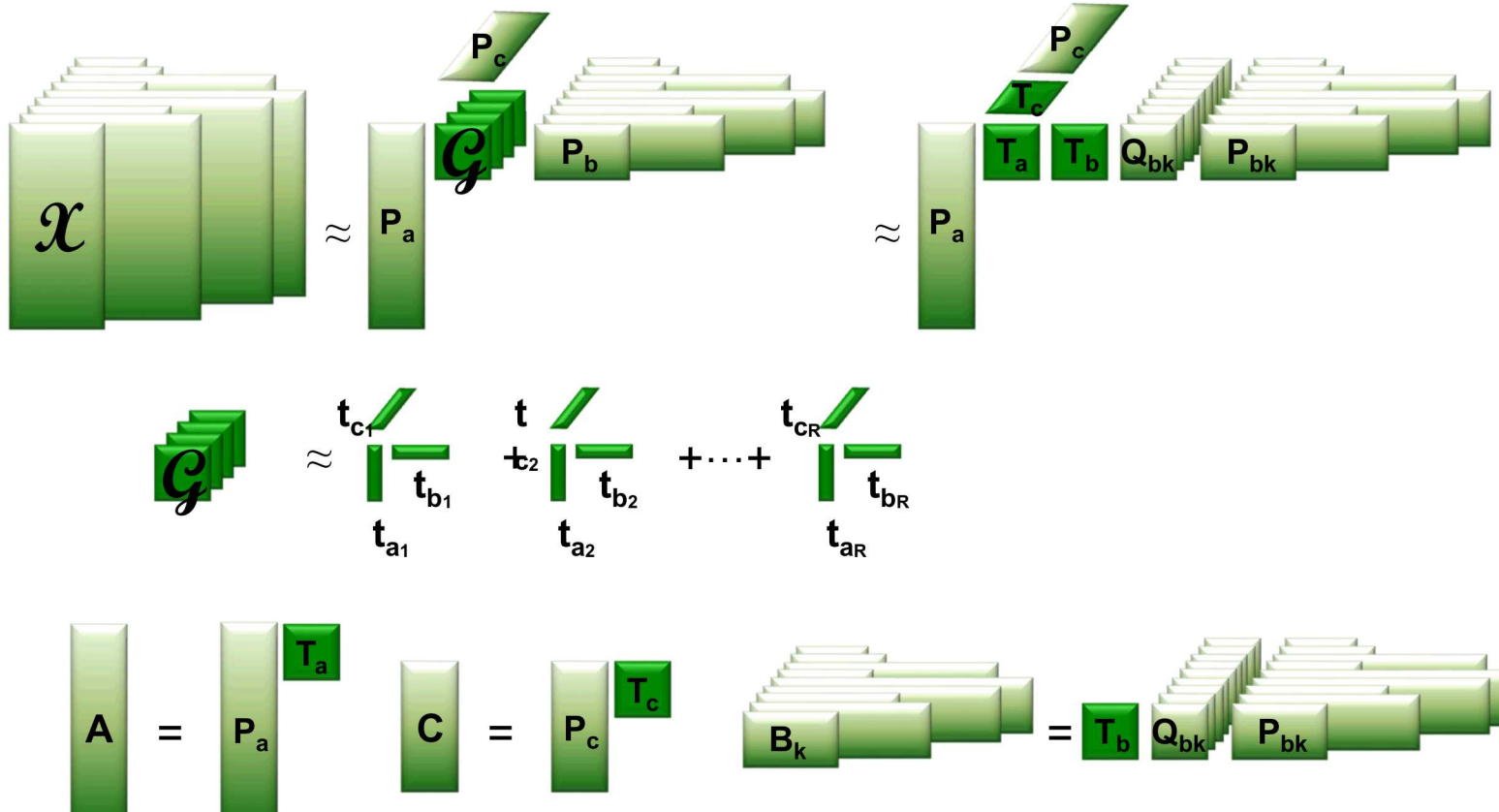
Tensor Factorization-PARAFAC2



- Data are initially individually factored
  - Orthogonal factors are retained and updated from PARAFAC on factored data



Tensor Factorization-Core PARAFAC2



PARAFAC2 Algorithms

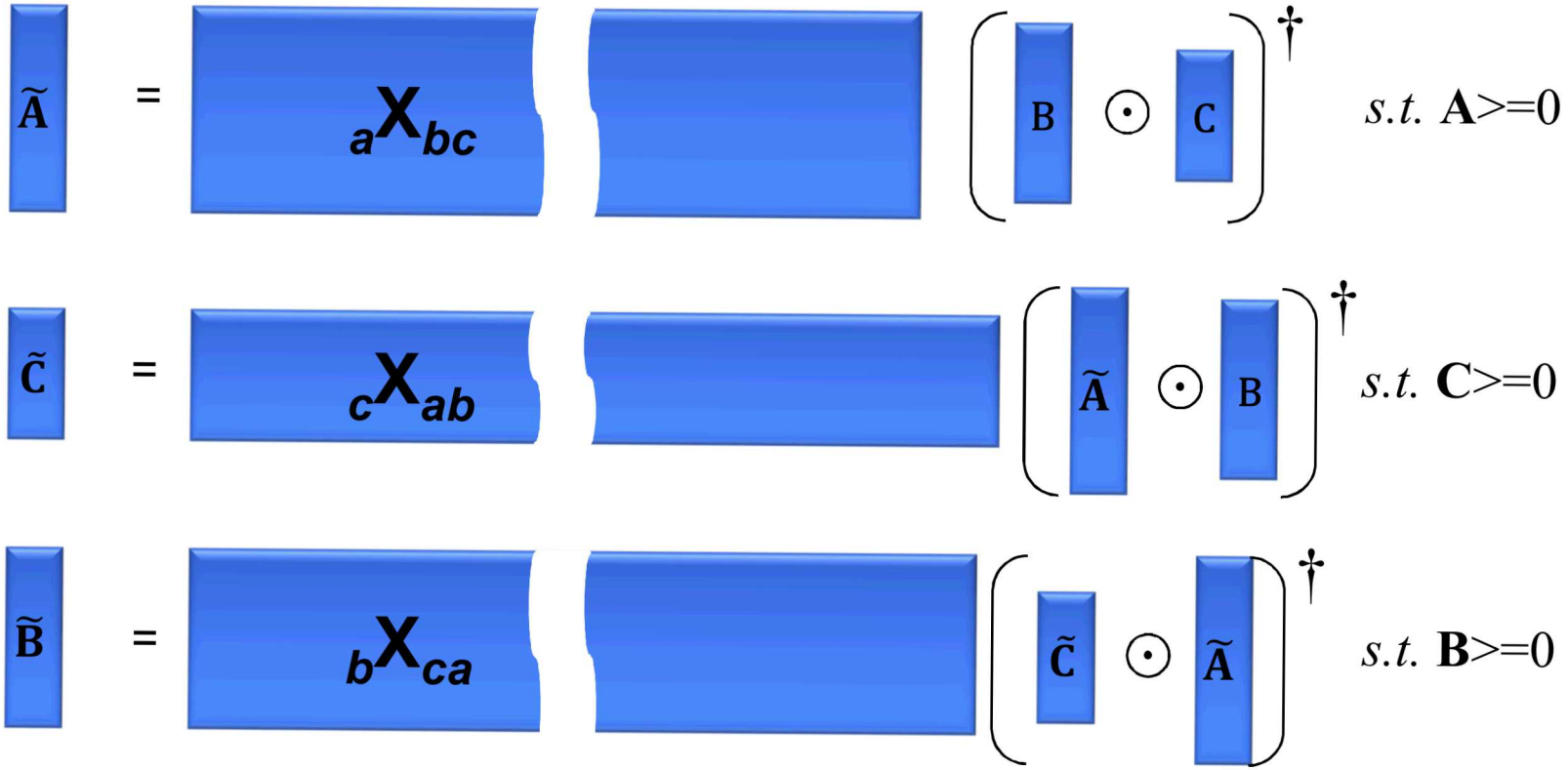
PARAFAC2

- Step 1: Initialize **A** as the eigenanalysis solution of $\mathbf{A}\Lambda = \left(\sum_k \mathbf{X}_k^{\mathrm{T}}\mathbf{X}\right)\mathbf{A}$. Initialize $\mathbf{T_b}$ and $C_1, \ldots, C_k$ as $\mathbf{I}_R$. Where the diagonal of C_k is the k^{th} column of **C**.
- Step 1a. Compute the SVD $\mathbf{T_b} C_k \mathbf{A}^{\mathrm{T}} \mathbf{X}_k^{\mathrm{T}} = \mathbf{U}_k \Delta_k \mathbf{V}_k^{\mathrm{T}}$ and update $\mathbf{P}_{\mathbf{b}k}$ as $\mathbf{V}_k \mathbf{U}_k^{\mathrm{T}}$.
- Step 1b. Update $\mathbf{T_b}$, **A** and $C_1, \ldots, C_k$ by PARAFAC algorithm applied to the three-way array **G**, with frontal planes $\mathbf{P}_{\mathbf{b}k}^{\mathrm{T}} \mathbf{X}_k$.
- Step 1c. Evaluate the residual function value $\sigma_3 = \sum_k \left\| \mathbf{X}_k - \mathbf{P}_{\mathbf{b}k} \mathbf{T_b} C_k \mathbf{A}^{\mathrm{T}} \right\|^2$. If $\sigma_3^{old} - \sigma_3^{new} > \varepsilon \sigma_3^{old}$ for some small value ε, go to Step 1a.

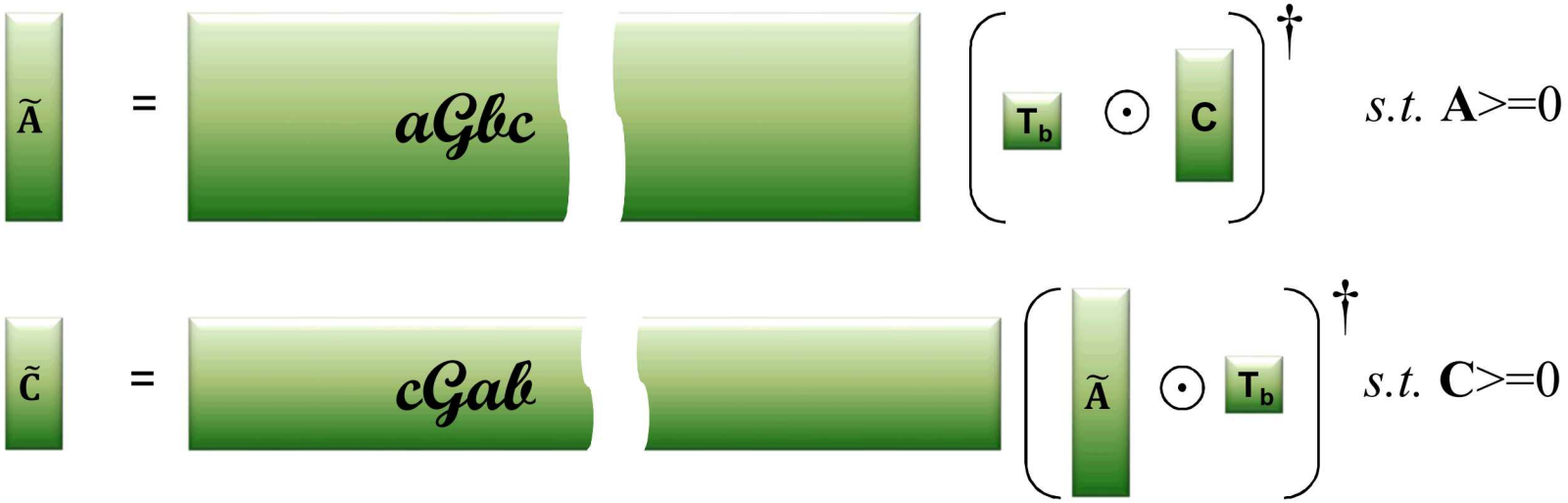
Core-PARAFAC2

- Step 1: Compute $\mathbf{P_a}$ as the eigenanalysis solution of $\mathbf{P_a}\Lambda = \left(\sum_k \mathbf{X}_k^{\mathrm{T}}\mathbf{X}\right)\mathbf{P_a}$. Initialize $\mathbf{T_a}$ and $\mathbf{T_b}$ as $\mathbf{I}_R$, and $C_1, \ldots, C_k$ as $\mathbf{I}_R$ plus $\mathbf{N}_R$. The diagonal of C_k is the k^{th} column of **C**.
- Step 1a. Compute the SVD $\mathbf{P}_{\mathrm{a}}^{\mathrm{T}} \mathbf{X}_k^{\mathrm{T}} = \mathbf{U}_k \Delta_k \mathbf{V}_k^{\mathrm{T}}$ and update $\mathbf{P}_{\mathbf{b}k}$ as $\mathbf{V}_k$, and let $\mathbf{Y}_k = \mathbf{U}_k \Delta_k$
- Step 1b. Compute the SVD $\mathbf{T_b} C_k \mathbf{T}_{\mathrm{a}}^{\mathrm{T}} \mathbf{Y}_k^{\mathrm{T}} = \mathbf{U}_k \Delta_k \mathbf{V}_k^{\mathrm{T}}$, update $\mathbf{Q}_{\mathbf{b}k}$ as $\mathbf{V}_k \mathbf{U}_k^{\mathrm{T}}$, and let $\mathbf{Z}_k = \mathbf{Y}_k \mathbf{Q}_{\mathbf{b}k}$.
- Step 1c. Compute the SVD $[\mathbf{Z}_1(:) \quad \cdots \quad \mathbf{Z}_k(:)] = \mathbf{U}\Delta\mathbf{V}$ and update $\mathbf{P_c}$ as **V** and update three-way array **G** as $\mathbf{U}\Delta$. Set $\mathbf{T_c} = \mathbf{P}_{\mathrm{c}}^{\mathrm{T}} \mathbf{C}$.
- Step 1d. Update $\mathbf{T_a}$, the $\mathbf{T_b}$, and $\mathbf{T_c}$ by one cycle of a PARAFAC algorithm applied to **G**.
- Step 1e. Evaluate the residual function value $\sigma_3 = \left\| \mathbf{G} - \otimes\left(\mathbf{T_a}, \mathbf{T_b}, \mathbf{T_c}\right) \right\|^2$. If $\sigma_3^{old} - \sigma_3^{new} > \varepsilon \sigma_3^{old}$ for some small value ε, go to Step 1a.

Nonnegativity-PARAFAC

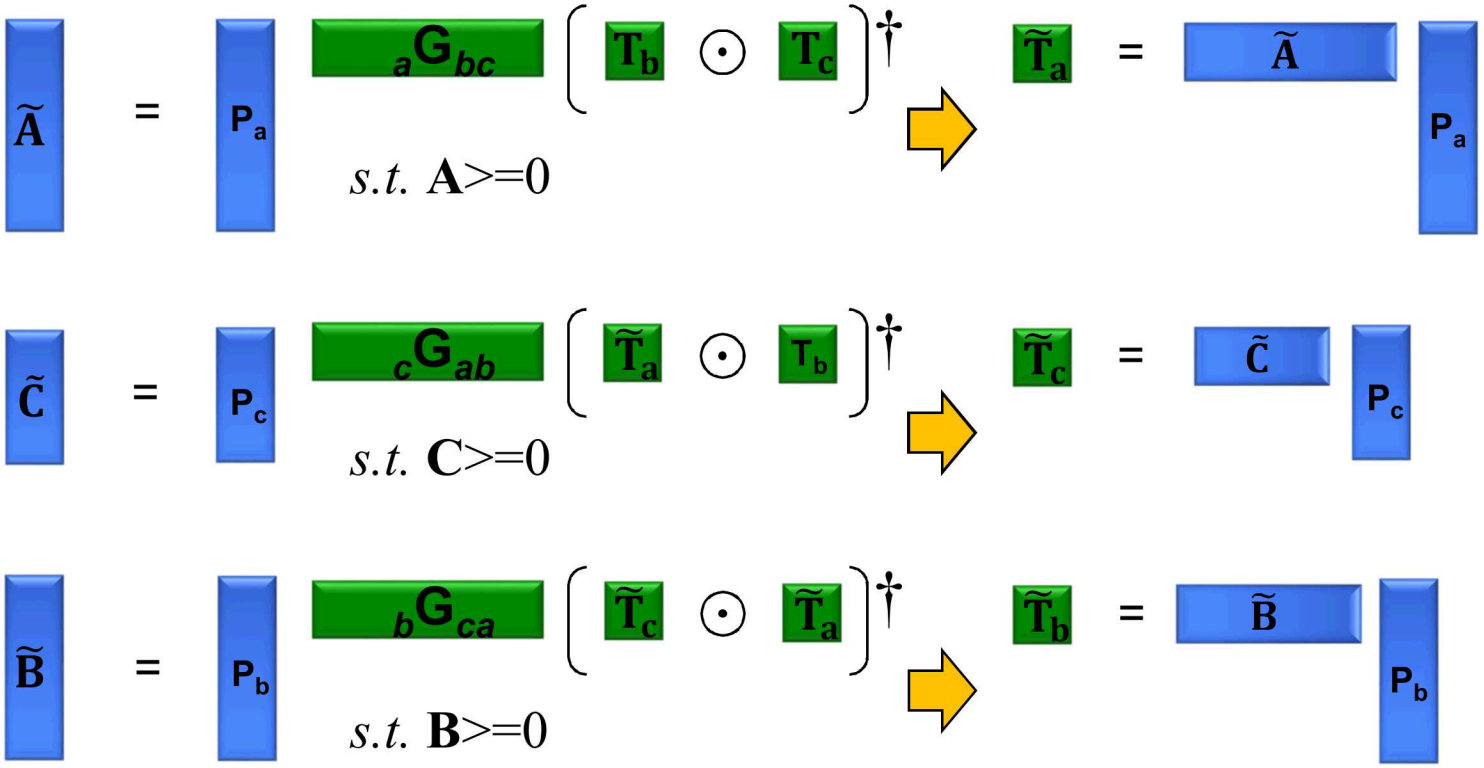


Nonnegativity-PARAFAC2



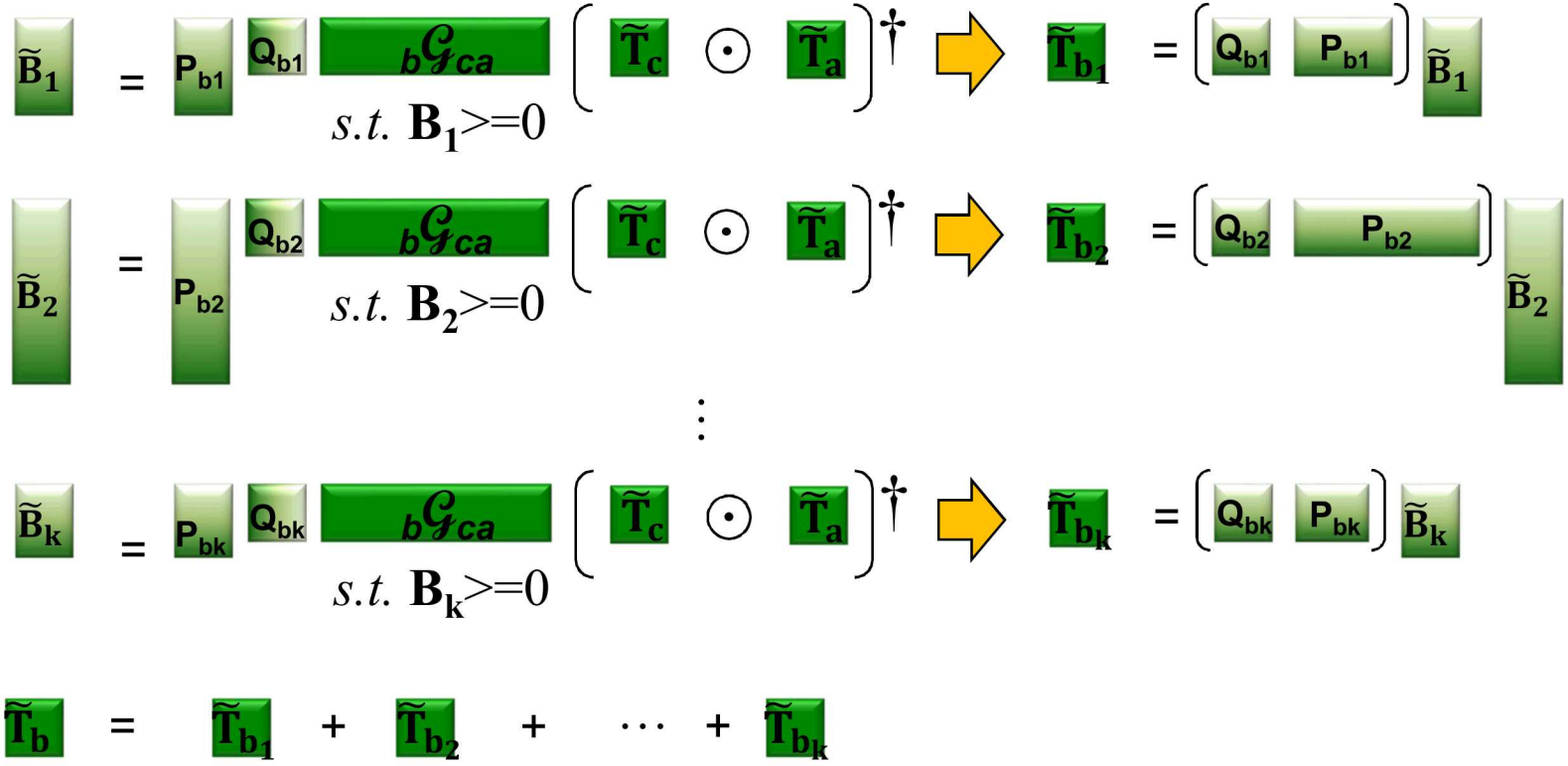
No mechanism for B-mode nonnegativity.
(New Flexible PARAFAC2 describes coupling method)

Nonnegativity-Core PARAFAC



1) Solve for the factors in a rigorous NNLS sense
2) Project nonnegative factors into orthogonal factor space
3) Yields rotation matrix for ~nonnegative factors

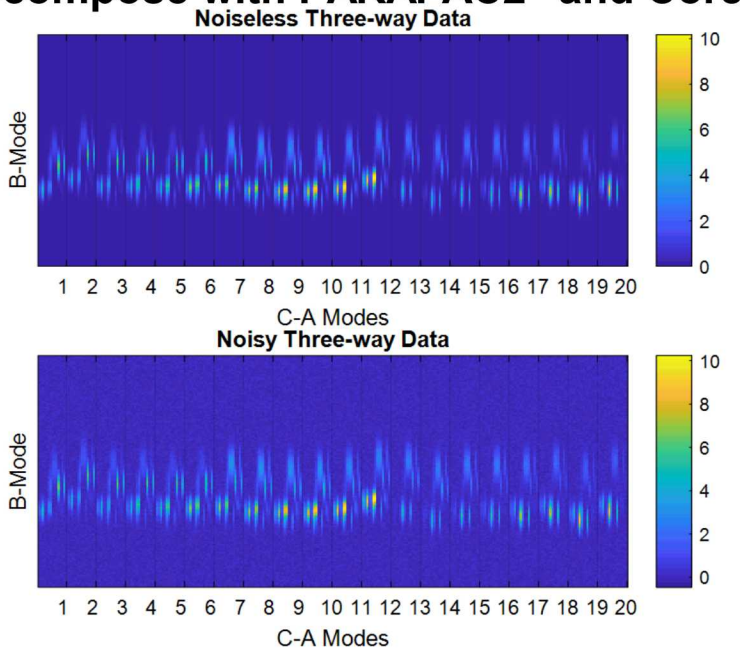
Nonnegativity-Core PARAFAC2

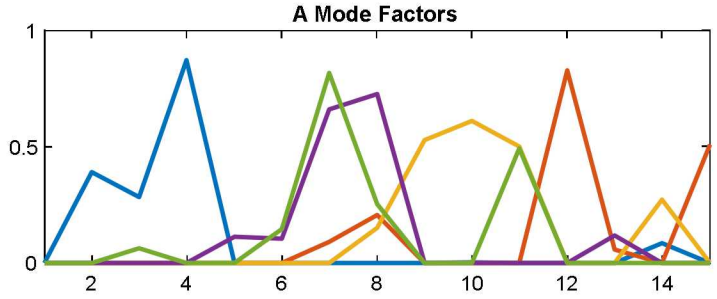


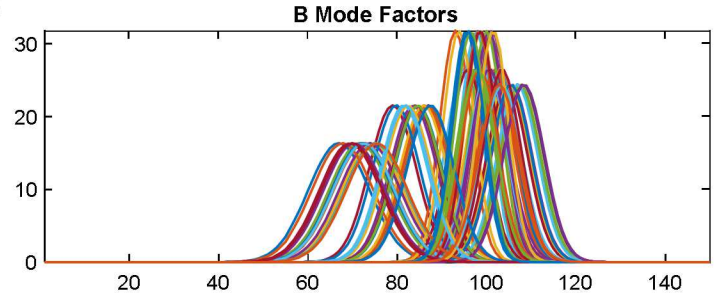
1) Solve for A and C modes rotation matrices as in NN-Core PARAFAC
2) For B-mode, solve for each layer factors in a rigorous NNLS sense
2) Project nonnegative factors into orthogonal factor space for each layer
3) Add rotation matrices for each layer to get overall B-mode rotation matrix

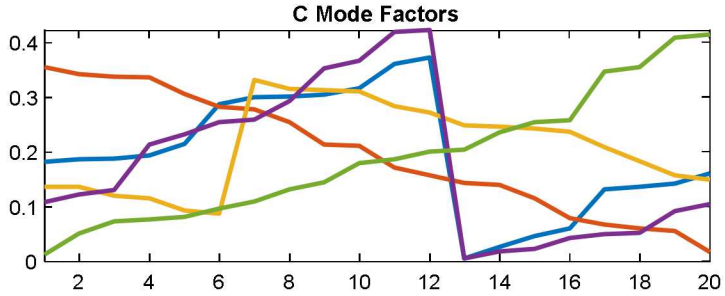
Synthetic Data

- **Data parameters:**
  - **Dimensions: 15 x 150 x 20 (A x B x C)**
  - **Rank 5**
  - **Shifts in Gaussian-shaped B-mode factors**
- **Gaussian Noise**
  - **Added to ~100 PSNR**
- **Decompose with PARAFAC2* and Core-PARAFAC2**





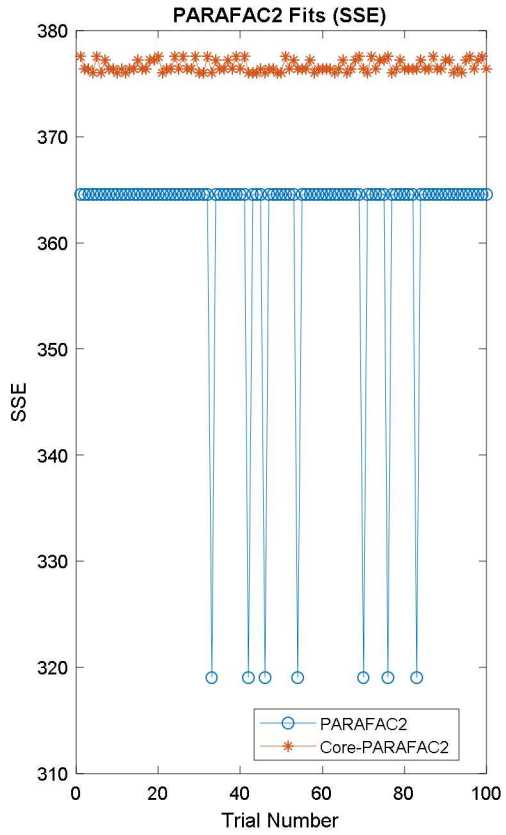


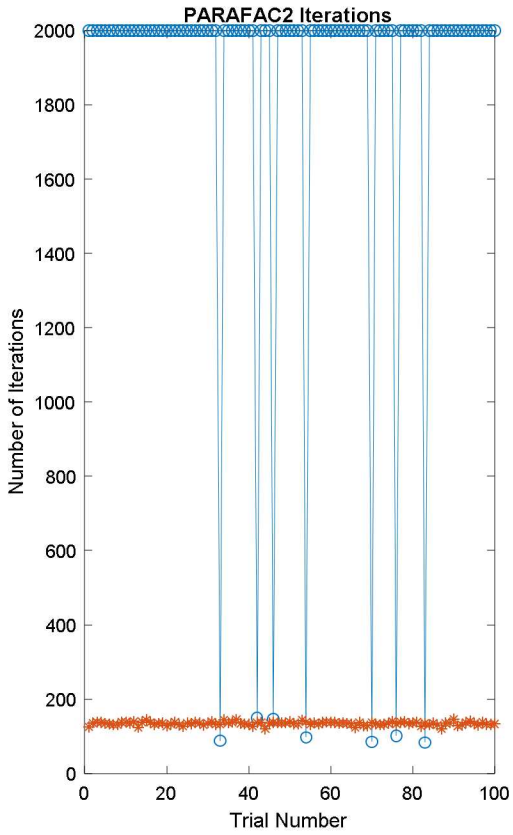


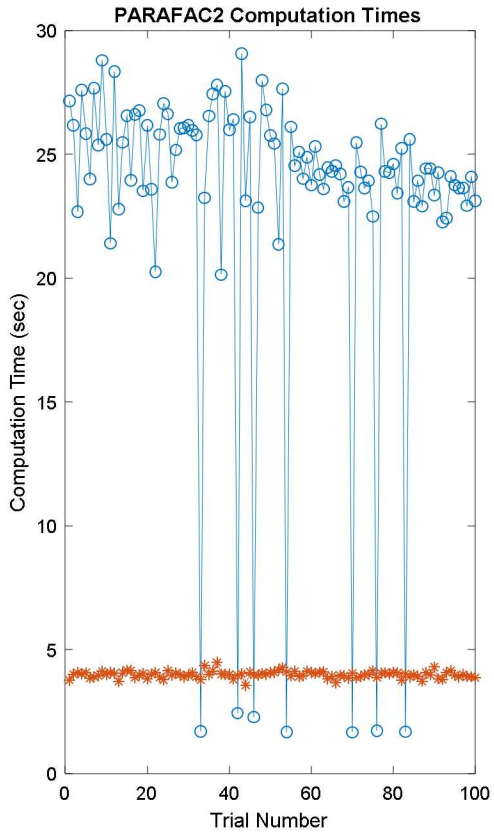
* https://www.mathworks.com/matlabcentral/mlc-downloads/downloads/submissions/1089/versions/1/download/zip/parafac2.zip (downloaded 5/5/18)

Sandia National Laboratories

PARAFA2 and Core PARAFAC2 Fitting Statistics



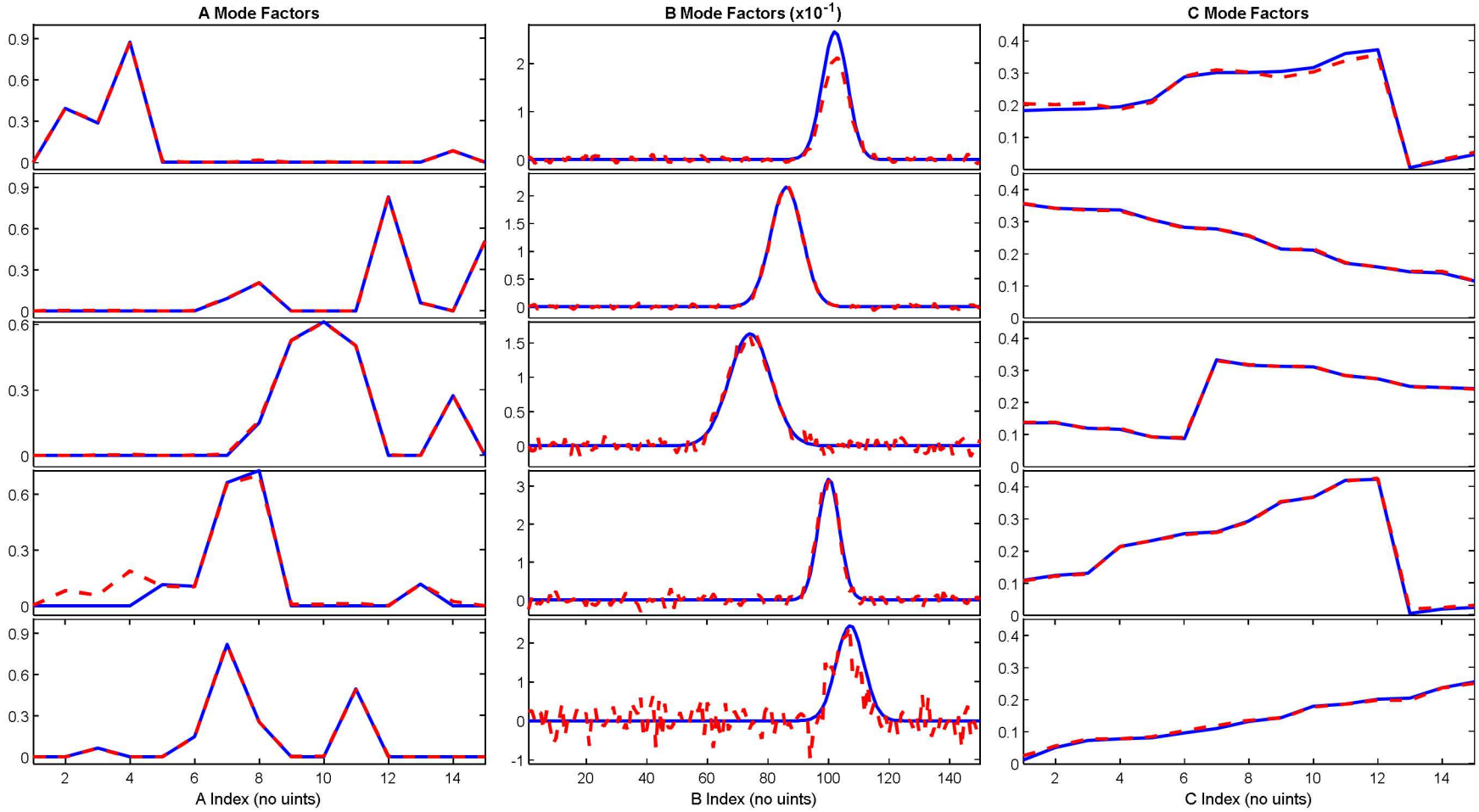




Core-PARAFAC2 appears to have much more computational stability
Standard PARAFAC2 produced "good" fits ~7% of the time.

Sandia National Laboratories

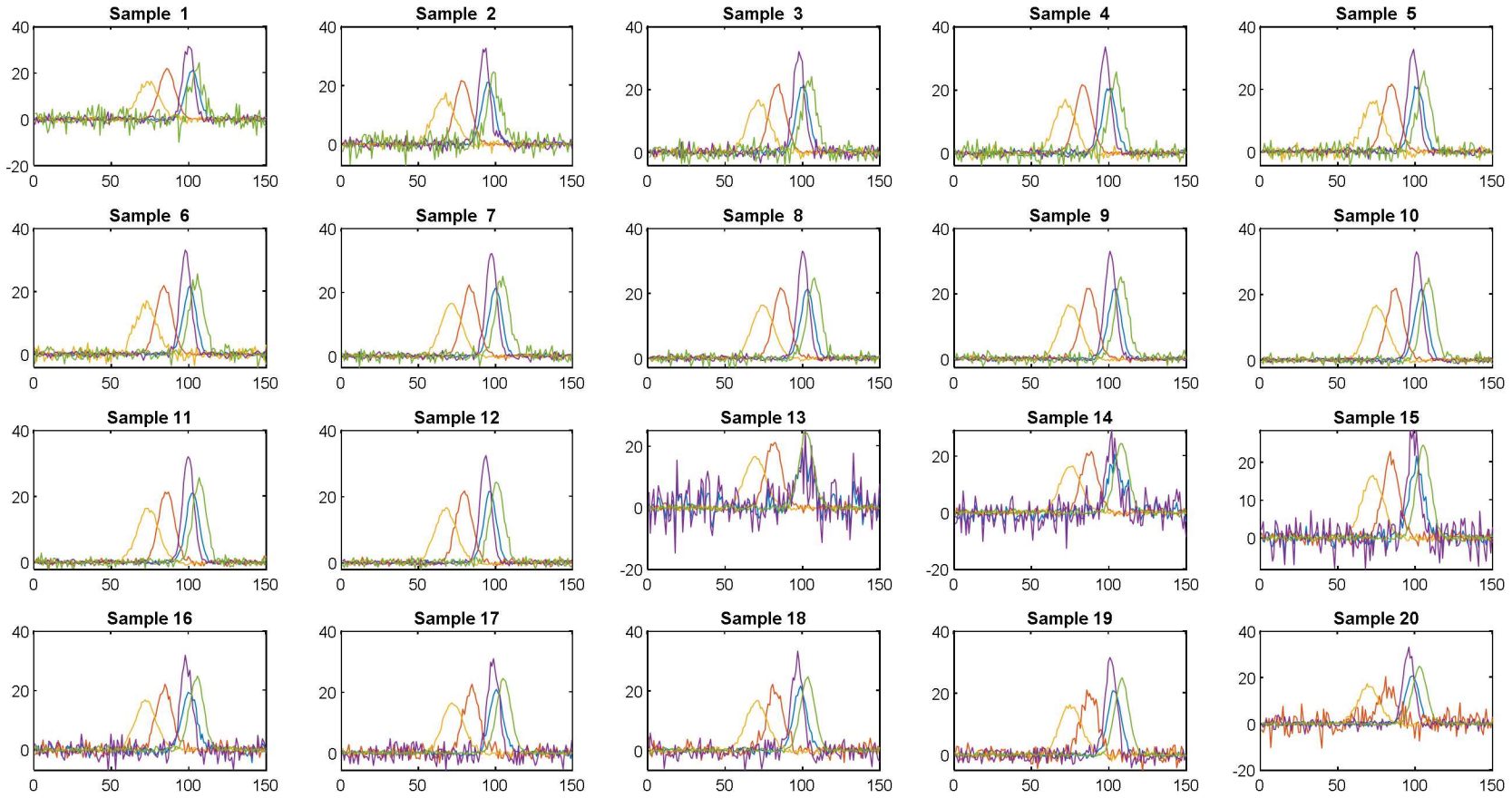
PARAFAC2 Best Fit (first slab)



PARAFAC2 does a nice job on this synthetic data set, some of the time.

Sandia National Laboratories

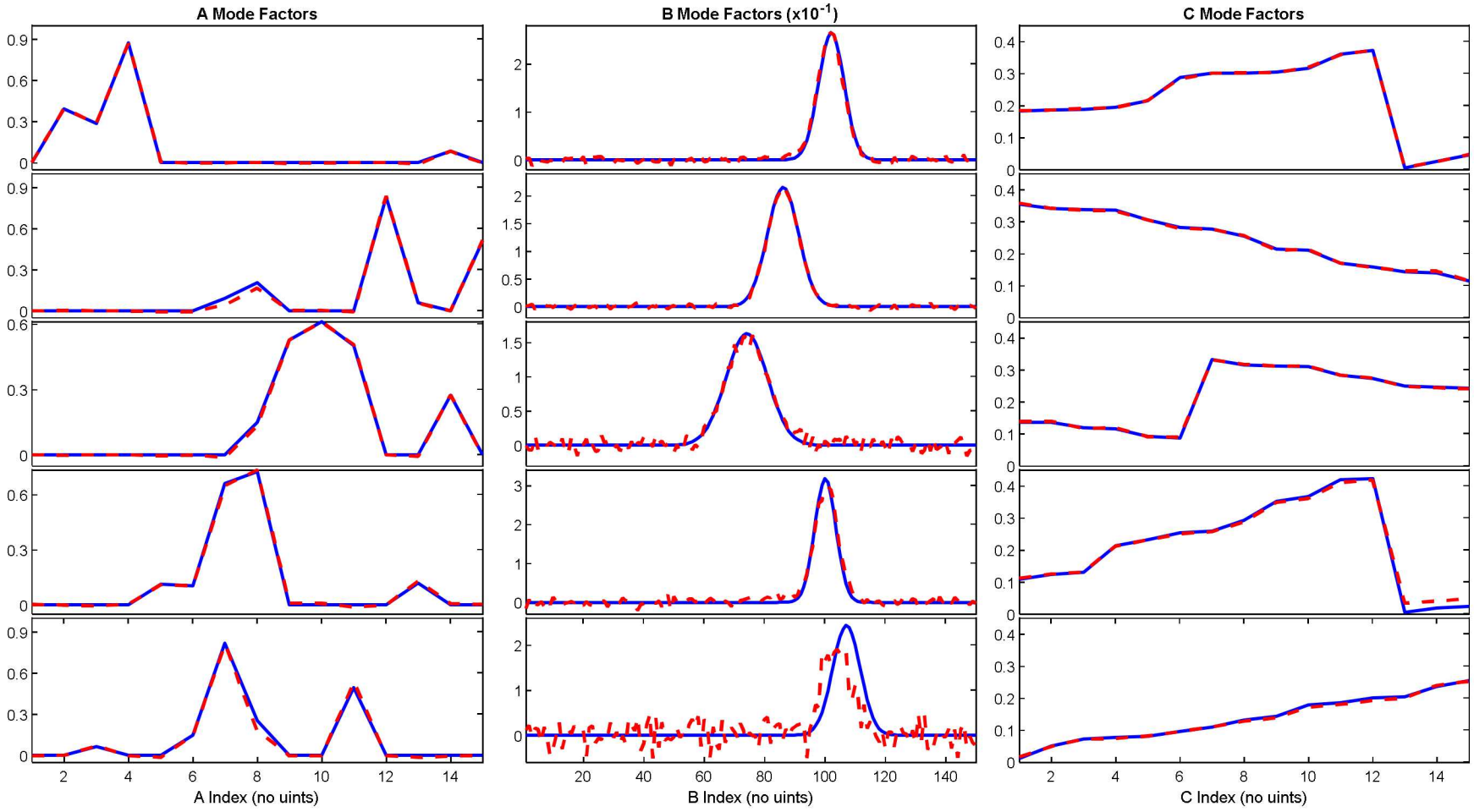
PARAFAC2 B-Mode Best Fit (all slabs)



Does not require nonnegativity constraints, but many trials.

Sandia National Laboratories

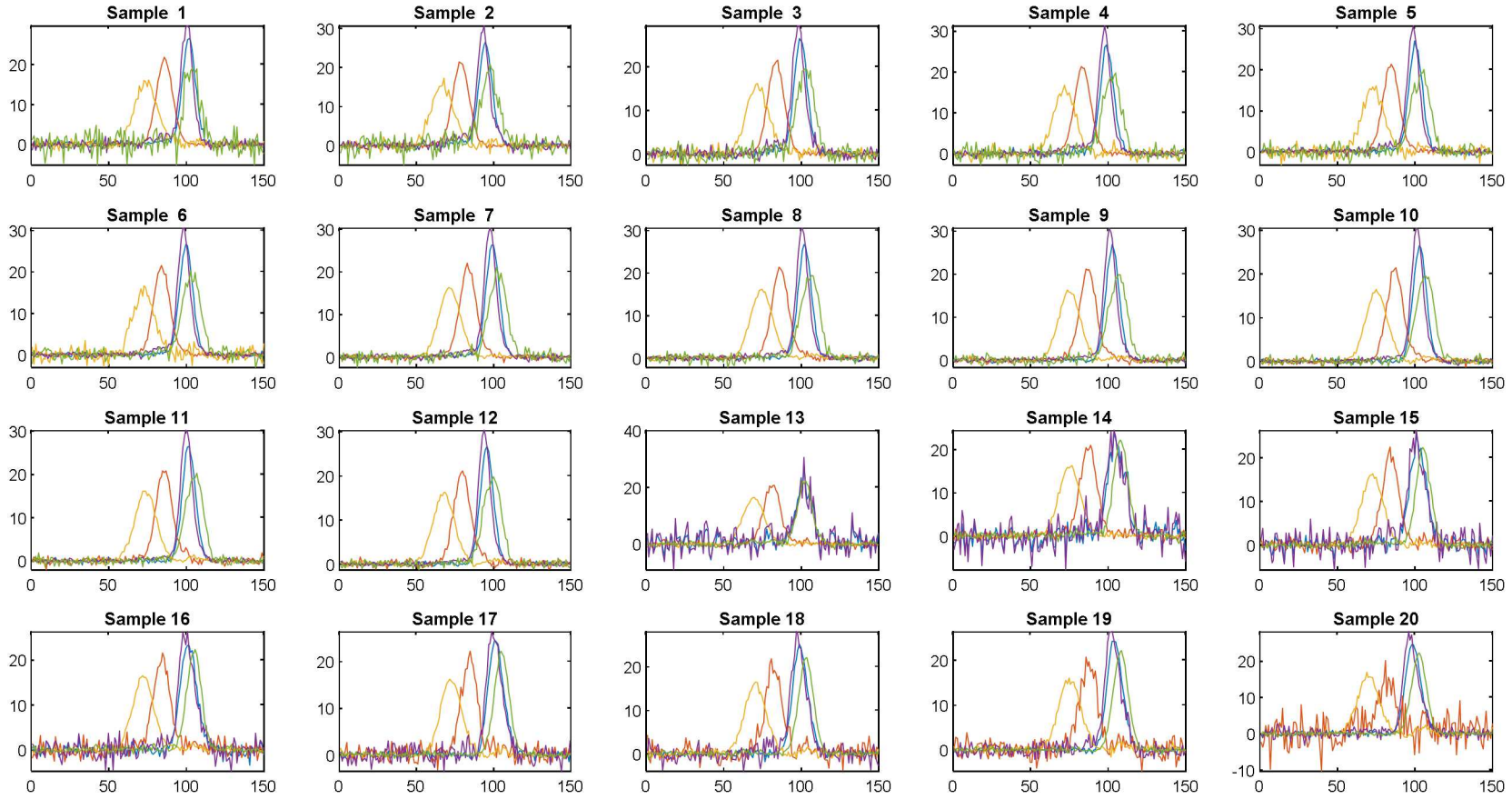
Core PARAFAC2 Best Fit (first slab)



Core PARAFAC2 also does a nice job on this synthetic data set.

Sandia National Laboratories

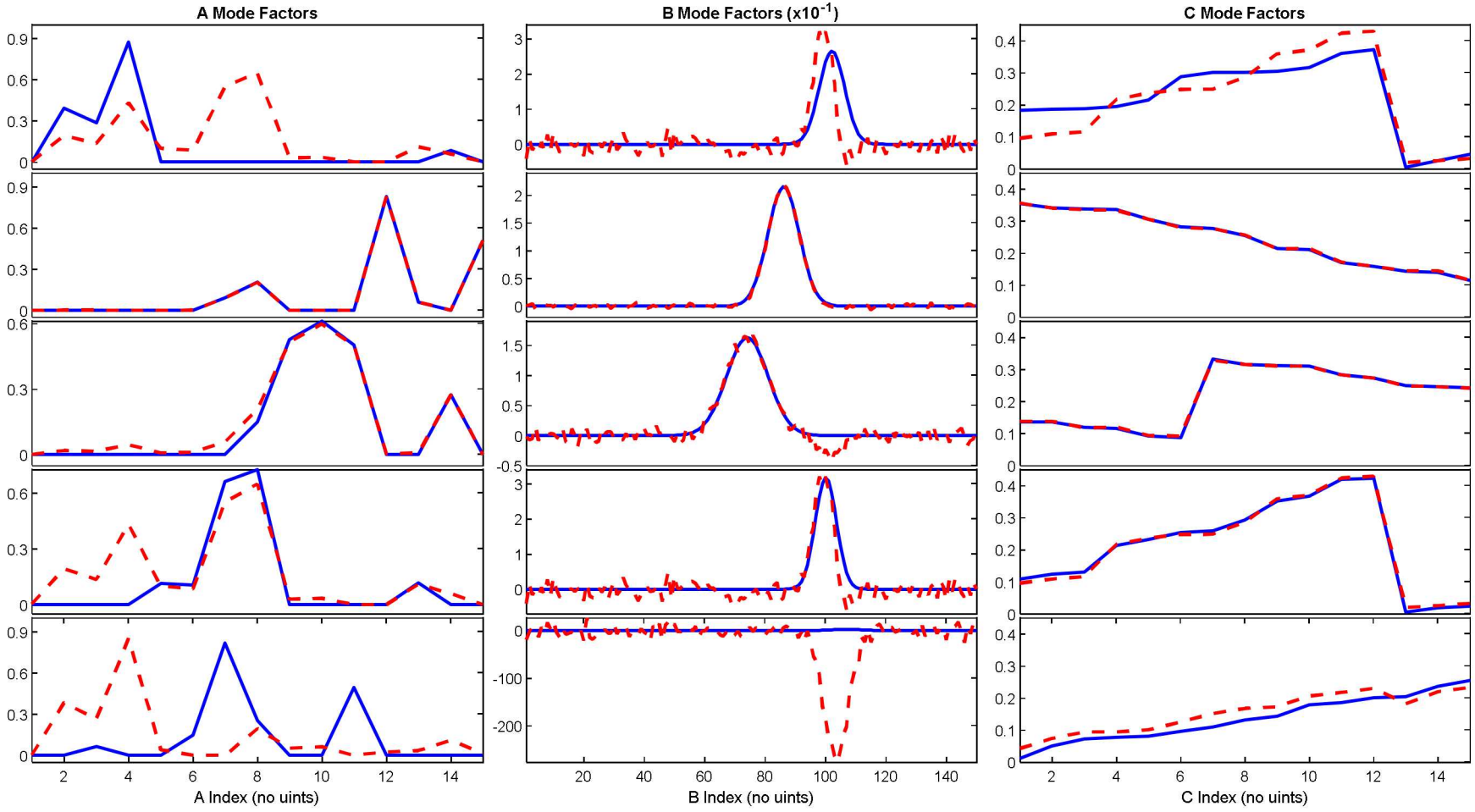
Core PARAFAC2 B-Mode Best Fit



Imposing nonnegativity constraints gives consistent results.

Sandia National Laboratories

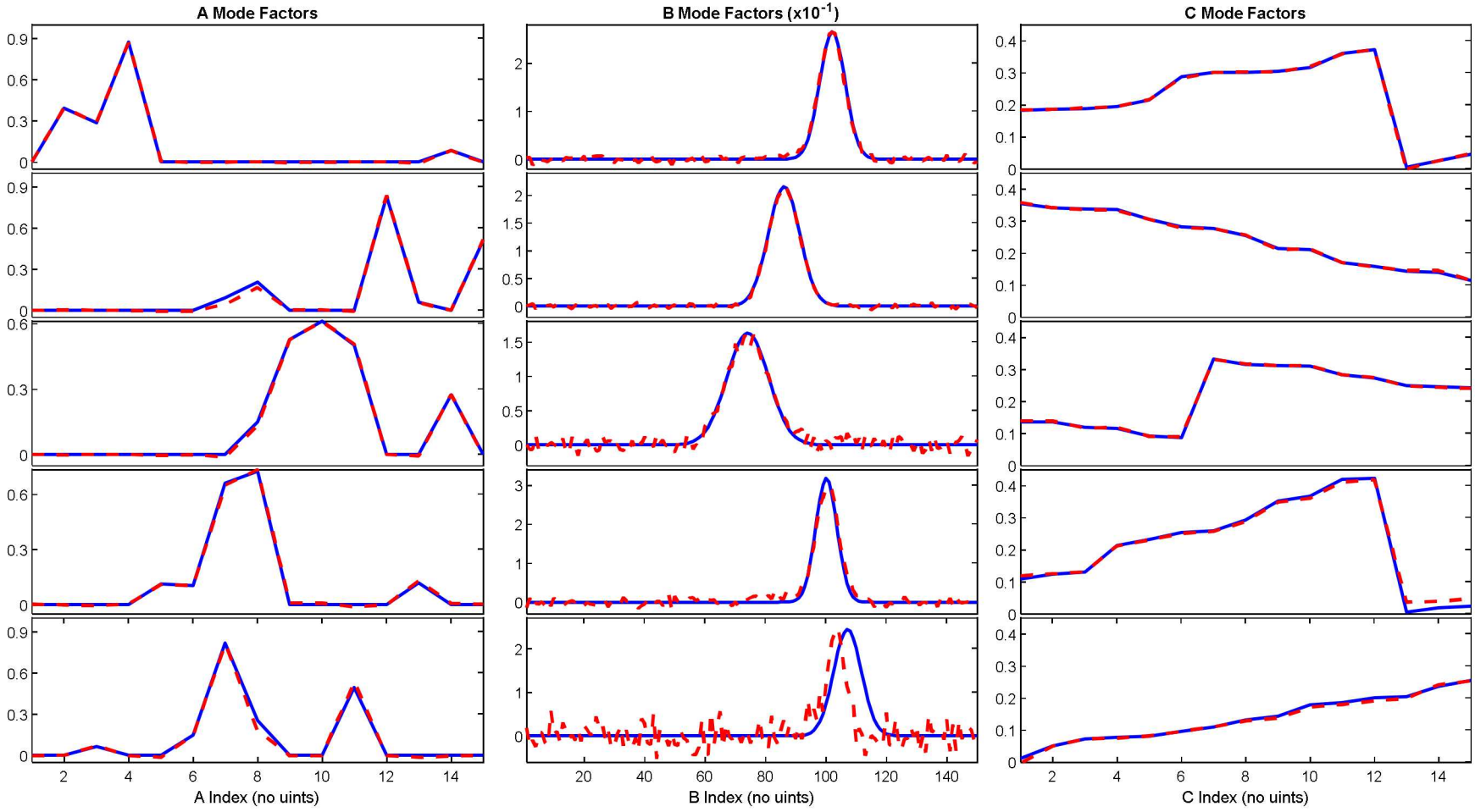
PARAFAC2 Worst Fit



PARAFAC2 sometimes produces suboptimal results.

Sandia National Laboratories

Core PARAFAC2 Worst Fit



Core PARAFAC2 with nonnegativity gives stable results.

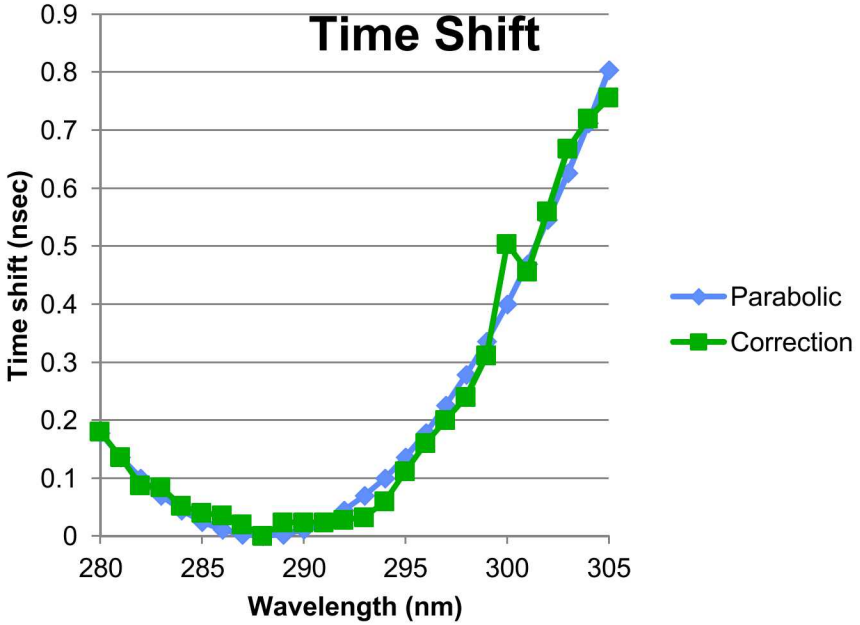
Sandia National Laboratories

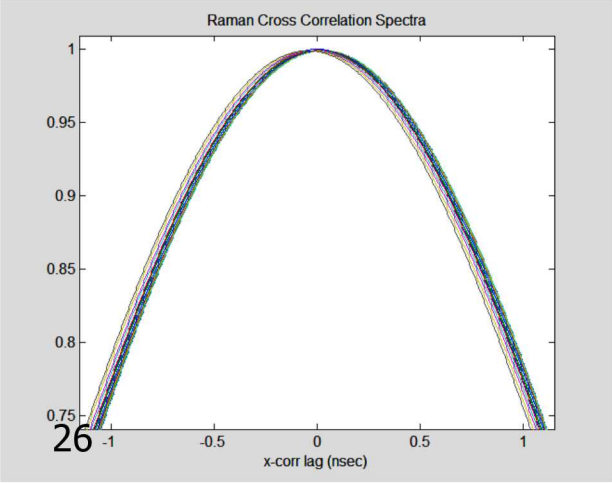
Experimental Data

- Solution of six EPA priority pollutants in methanol
  - Chemical Species:
    - Fluorene, Naphthalene, Pyrenene, Phenanthrene, Benzo[k]fluoranthrene, Benzo[b]fluoranthrene
    - From Sigma-Aldrich used without further purification
    - Solutions prepared in HPLC grade methanol.
- Collect a single TREEM
  - Original data size:
    - 640 time-mode elements
    - 251 emission wavelength-mode elements
    - 26 excitation wavelength-mode elements
- Acquisition
  - Collect waveforms: ~130 nsec duration in 200 psec increments
  - Step emission wavelengths: 300 – 550 nm in 1 nm increments
  - Excitation wavelengths: 280 – 305 nm in 1 nm increments
  - Collect each WTM in approximately 5 minutes

Correction of Time Domain Shifts

- Due to the laser gain profile, there is an excitation wavelength dependent shift in the time domain
  - Shift is approximately parabolic with wavelength
- Correction involves estimating shift with Raman or Rayleigh scattered excitation
  - Interpolate profiles at 50x measured data
  - Find shift values using cross-correlation
- Correct TREEM data
  - Interpolate profiles at 50x measured
  - Shift data requisite intervals
  - Down-sample data
  - Size after: 636 elements





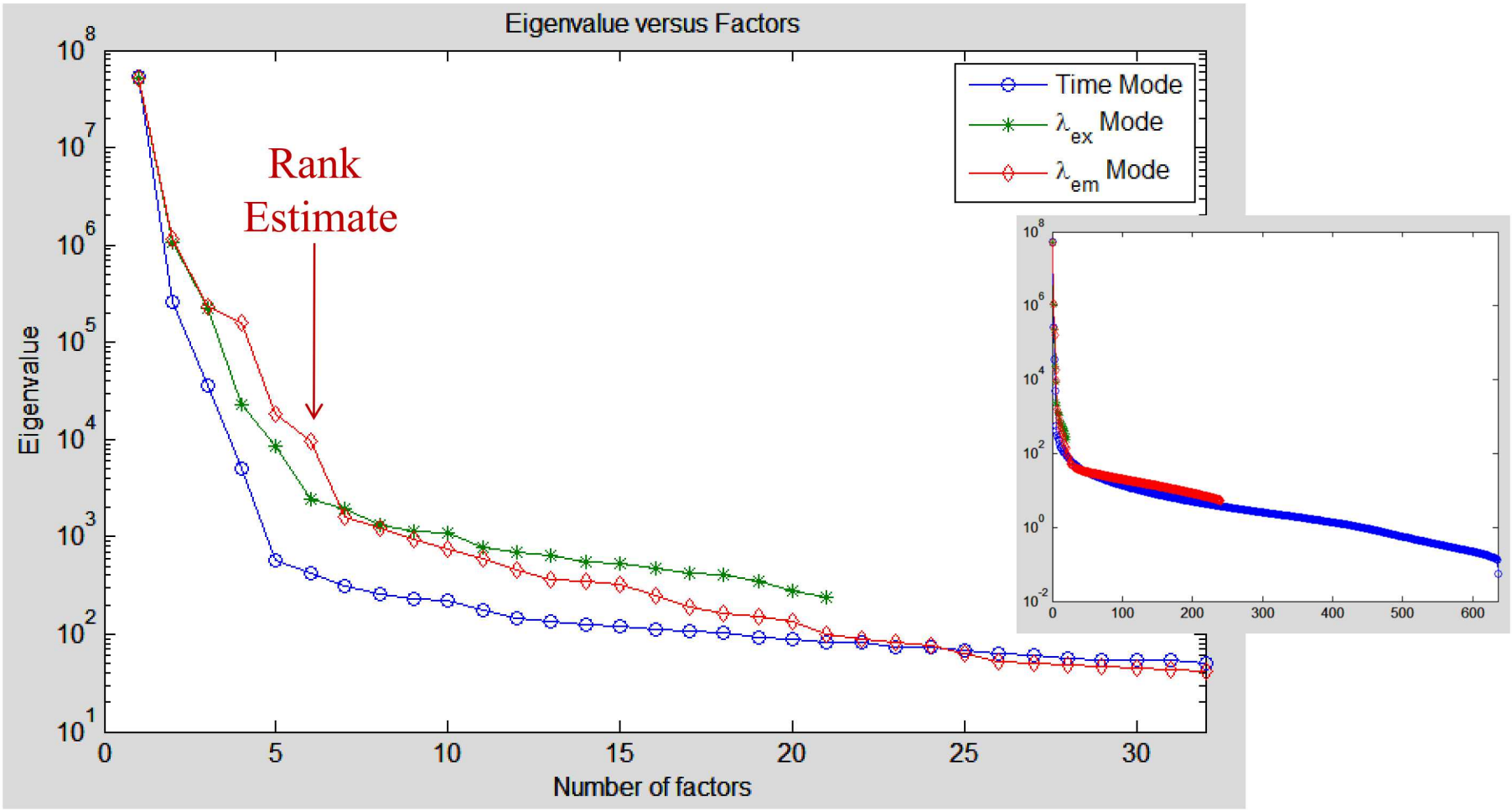
Sandia National Laboratories

Data Analysis

- Wavelength ranges, λ_{ex}: 280-300 nm, λ_{em}: 310-550 nm
- PARAFAC on shift-corrected data
  - Use fast Tucker1 based PARAFAC algorithm
  - Initialize with random factors, 6 factor model
  - Employ fast rigorous nonnegative least squares
  - Data size: 636×241×21
- PARAFAC2
  - Utilize code from Bro as published[1]
  - Initialize with default, "best of 10 runs…", 100 separate trials
  - Impose nonnegativity in excitation & emission modes
  - Data size 640×241×21
- Core-PARAFAC2
  - Utilize code developed at SNL
  - Initialize with random factor matrices, 100 separate trials
  - Impose nonnegativity in excitation, emission & temporal modes
  - Data size 640×241×21
- Computations were performed using MATLAB®
  - Dell Precision T7910 workstation, with two, 14-core, 2.6 GHz Intel® Xeon (ES-2697) processors and 192 Gbyte RAM

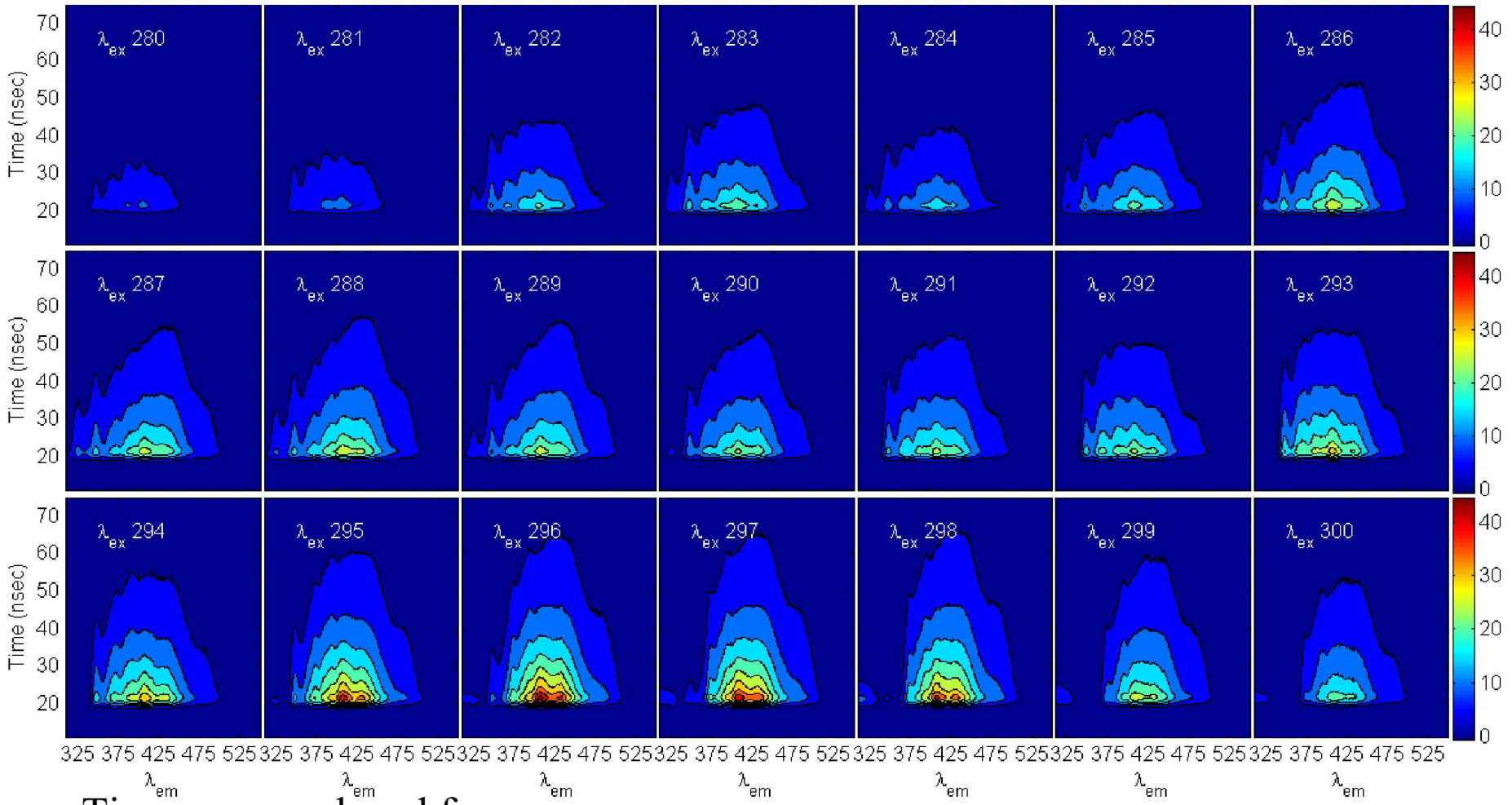
1. Kiers, H.A.L. TenBerge, J.M.F. Bro, R. J. Chemom., 13 (1999) 275-294.

Eigenvalue Plot



Sandia National Laboratories

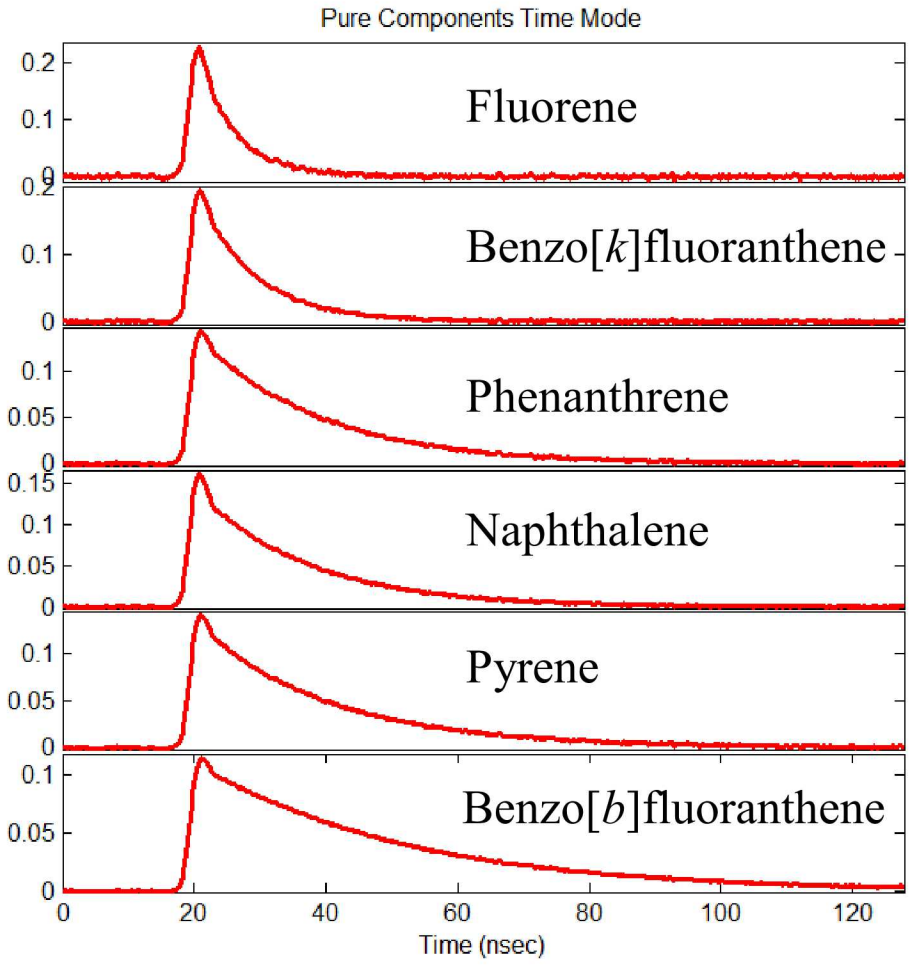
Raw TREEM Data as 21 WTMs

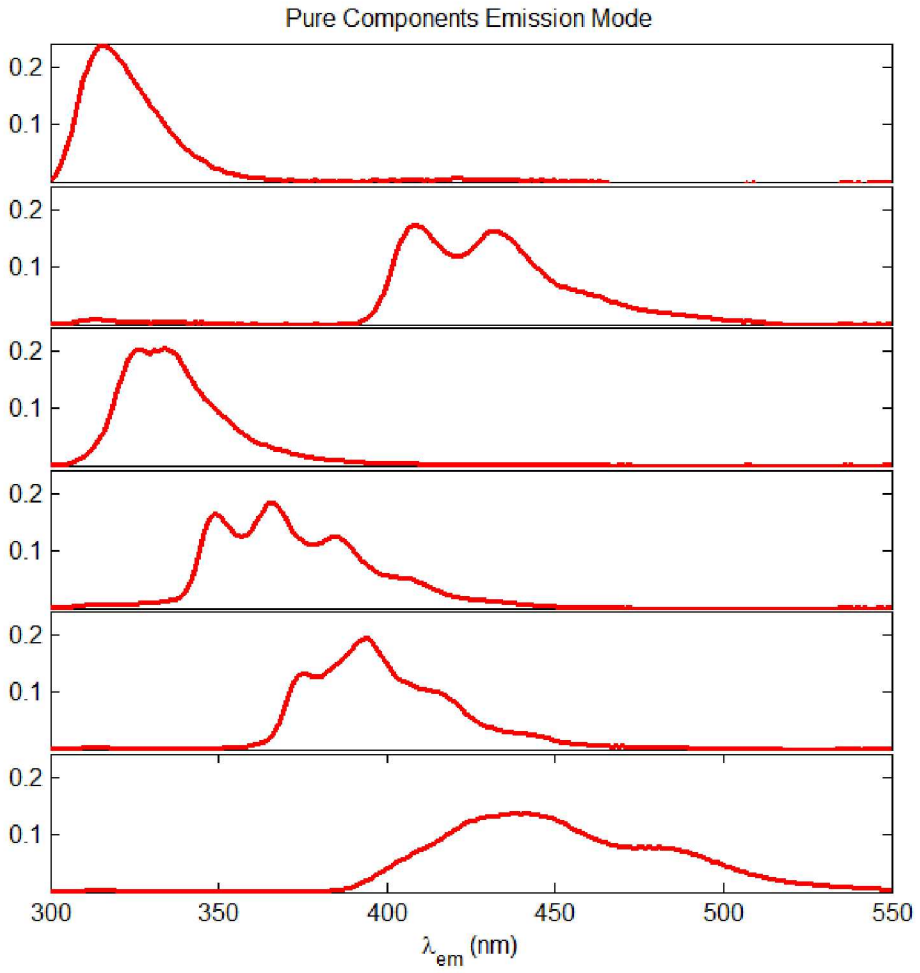


Time range reduced for improved comparison.

Sandia National Laboratories

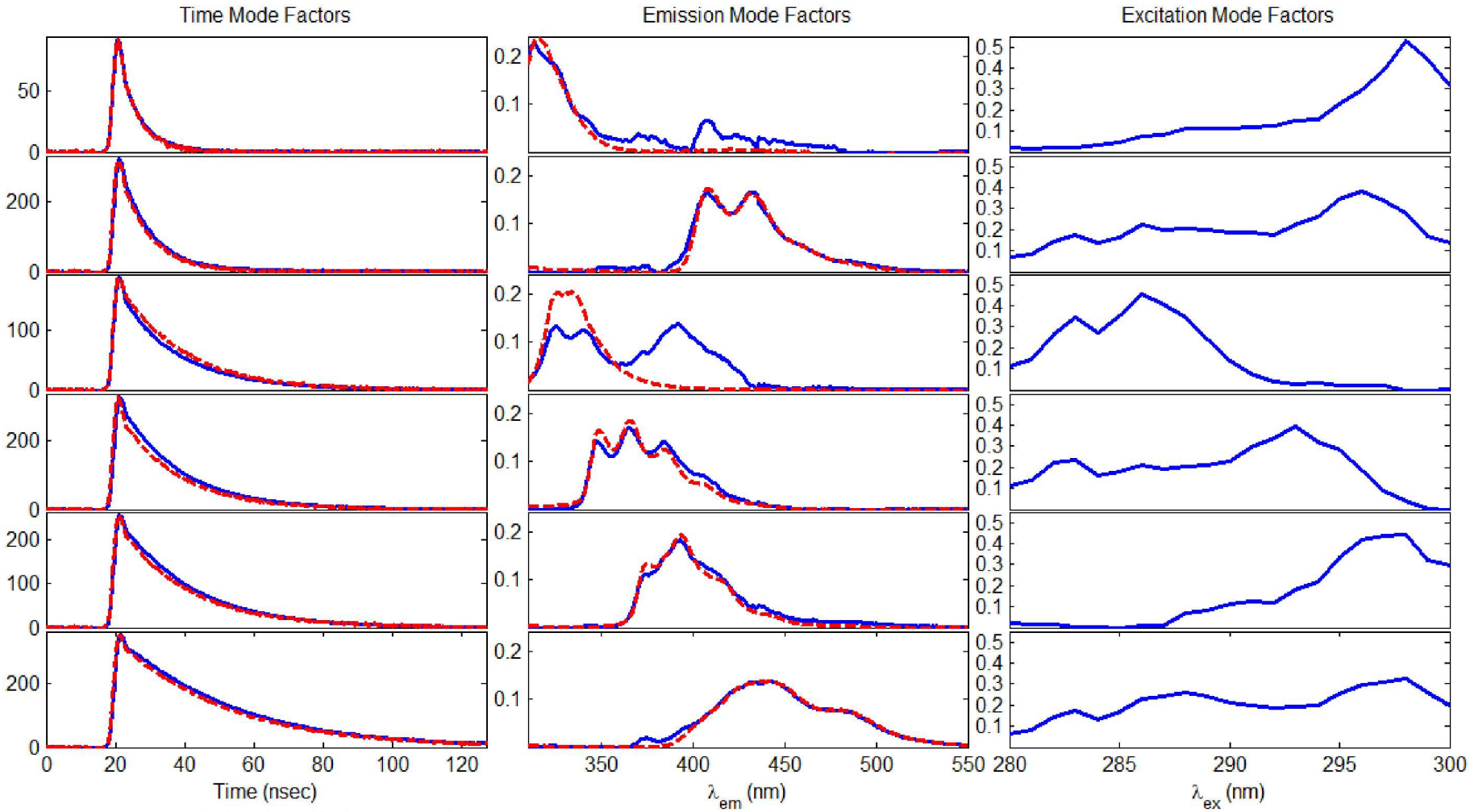
Pure Component Time and Emission Modes





Sandia National Laboratories

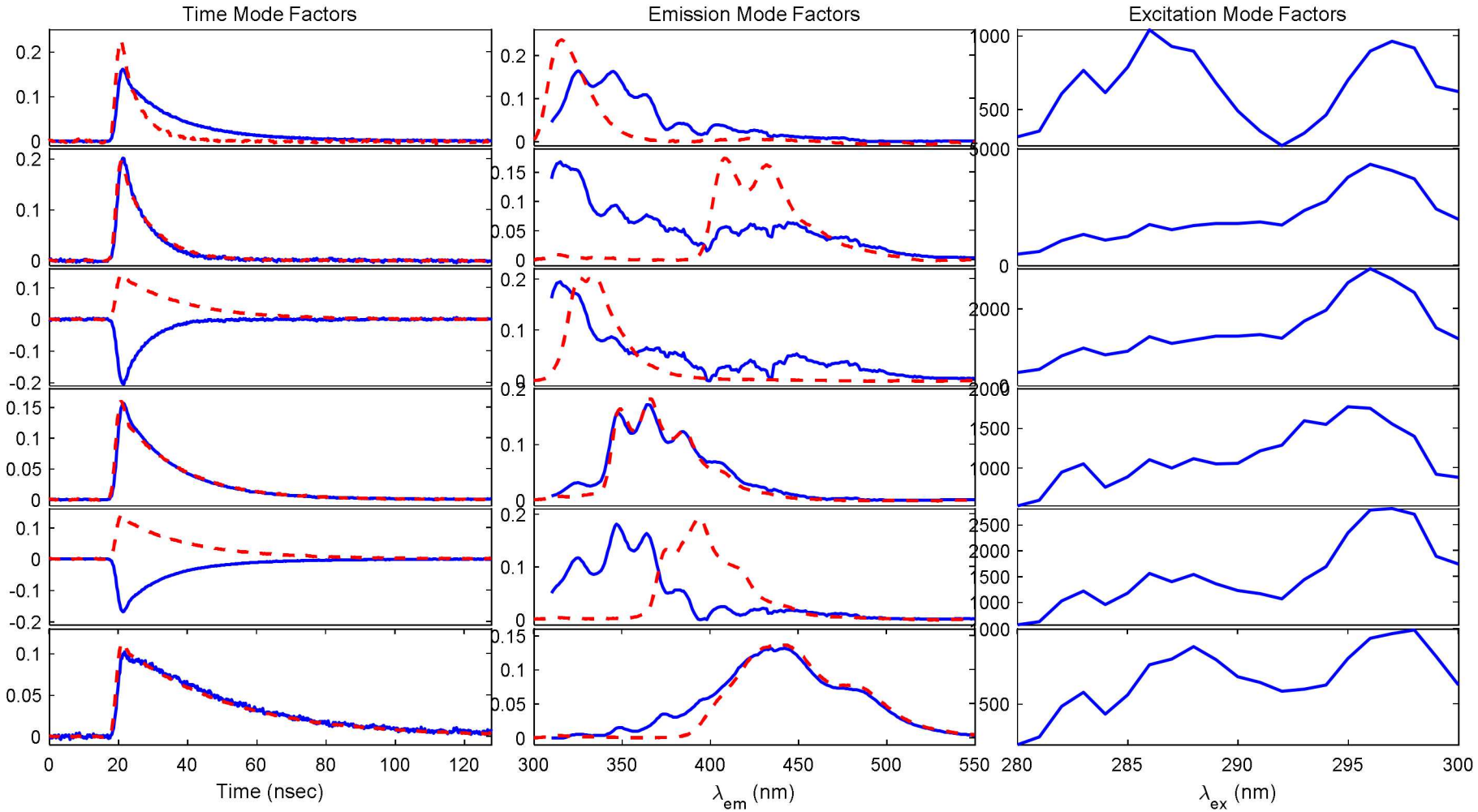
Six Factor Fast-PARAFAC Model



Nonnegativity in all three modes.
Computation time: 14 seconds
Blue: Fast PARAFAC, Red: Pure Components

Sandia National Laboratories

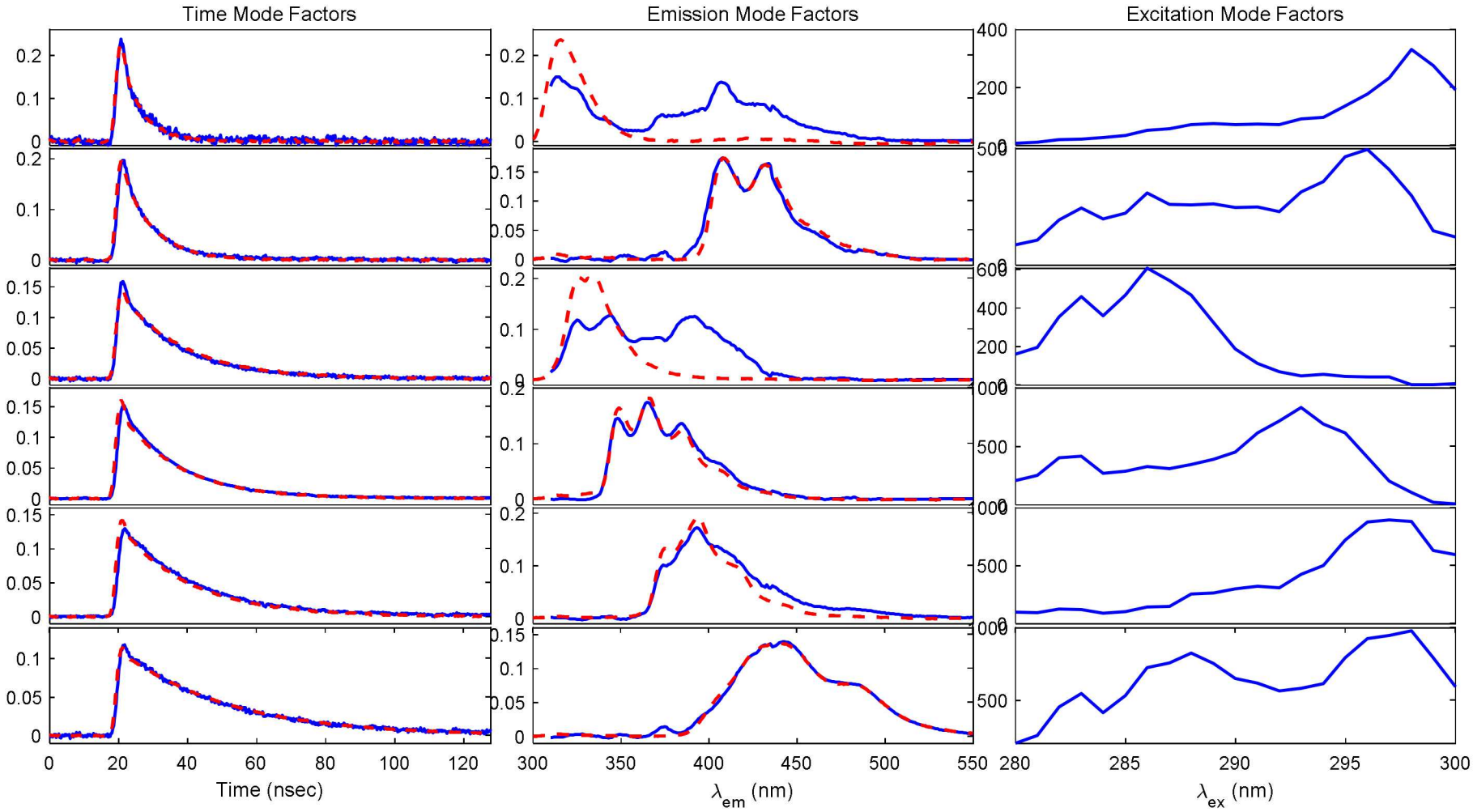
Six Factor PARAFAC2 Model Best Fit



Nonnegativity in Emission and Excitation modes.
Computation time: 357 seconds
Blue: PARAFAC2, Red: Pure Components

Sandia National Laboratories

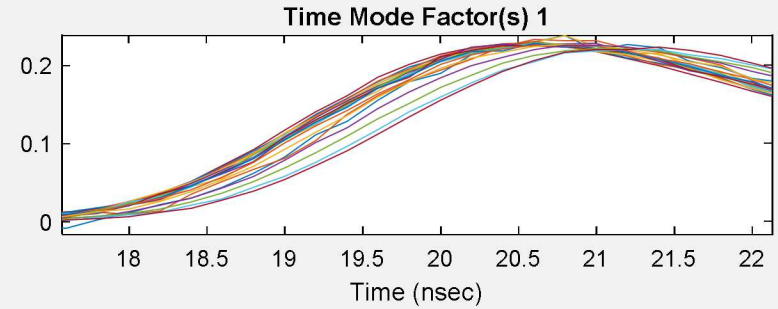
Six Factor Core-PARAFAC2 Model Best Fit

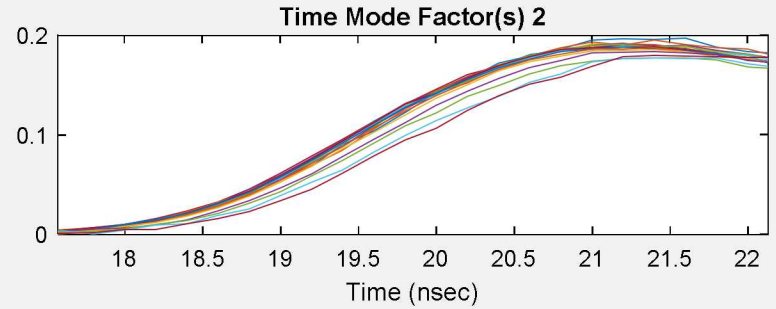


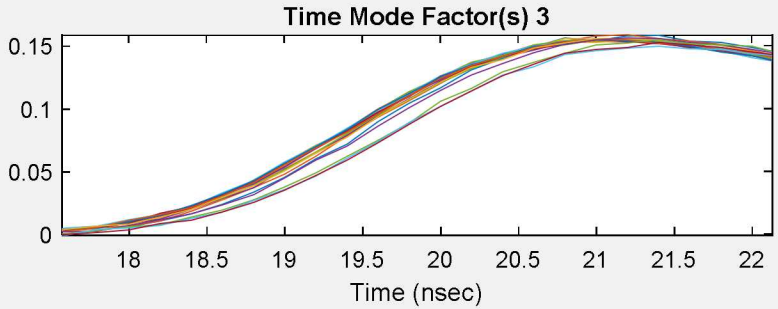
Nonnegativity in all three modes.
Computation time: 147 seconds
Blue: Core-PARAFAC2, Red: Pure Components

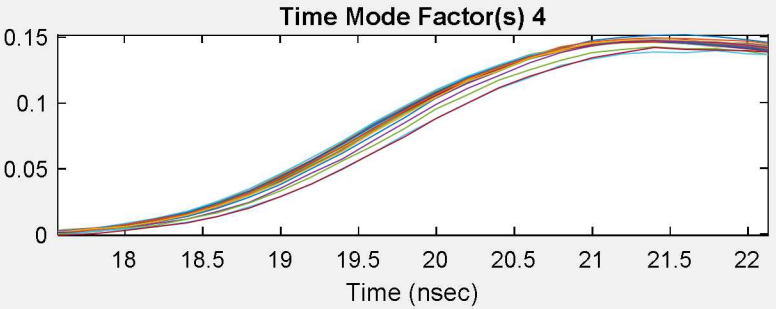
Sandia National Laboratories

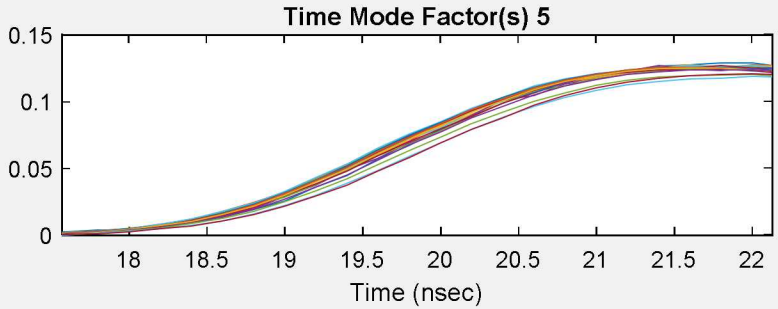
All Time (shifted) Mode Factors

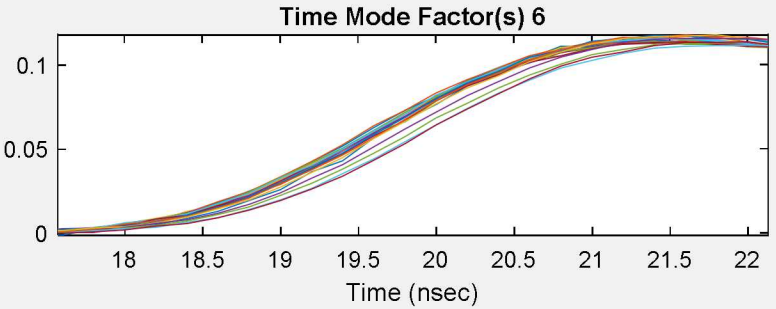




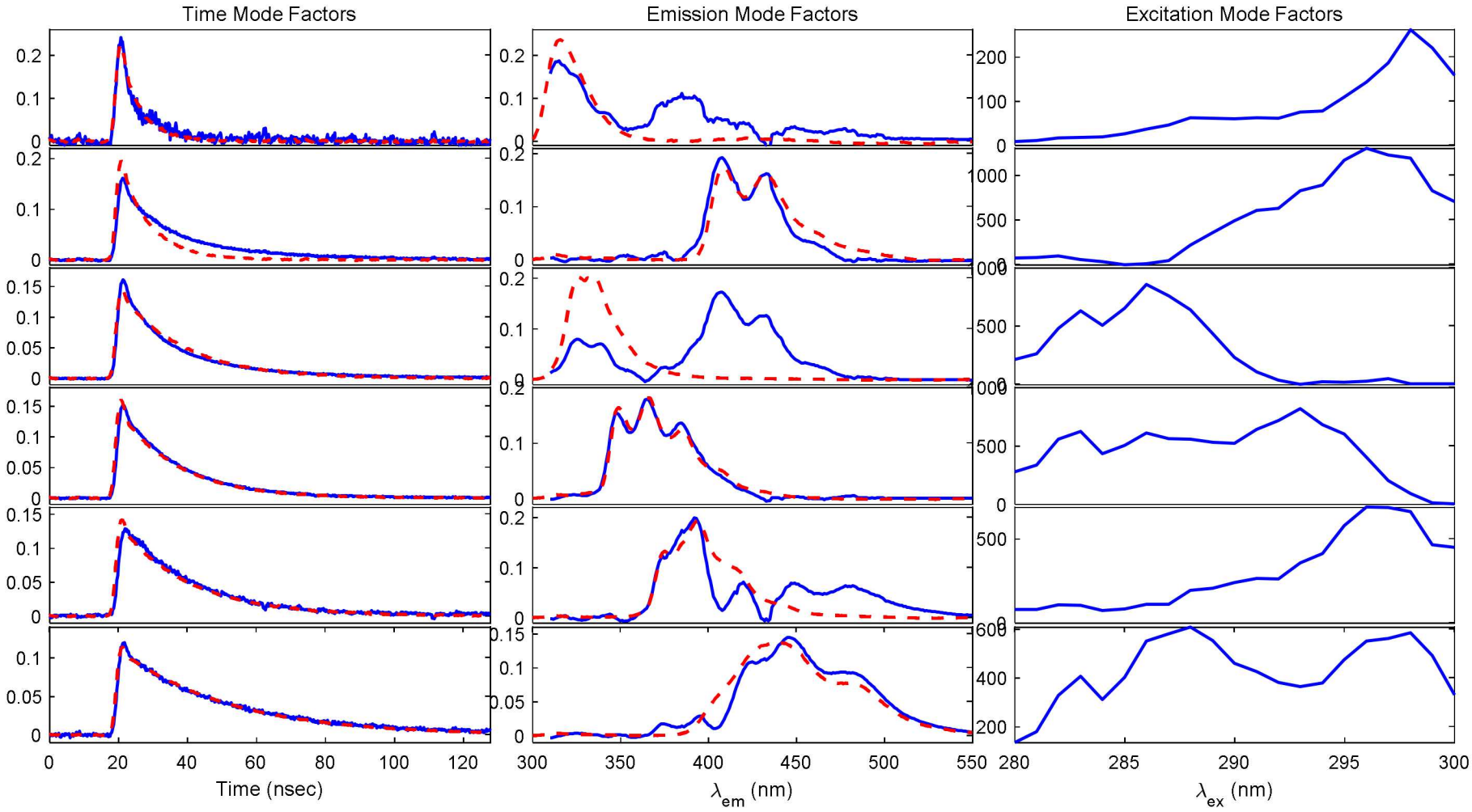






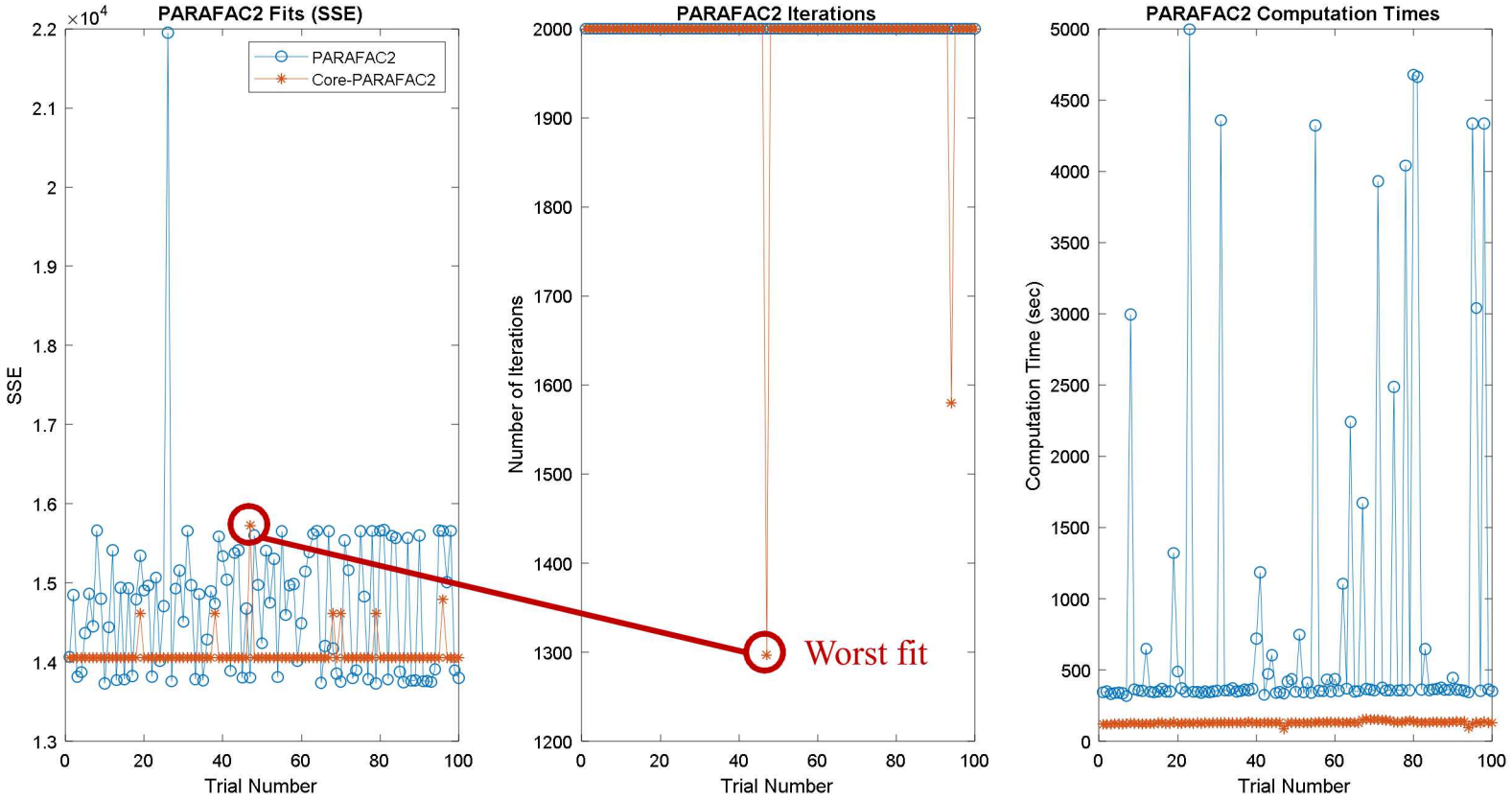


Six Factor Core-PARAFAC2 Model Worst Fit



Sandia National Laboratories

PARAFA2 and Core PARAFAC2 Fitting Statistics



Sandia National Laboratories

Results and Conclusions

- Core PARAFAC2 with nonnegativity, under appropriate circumstances, provides excellent, stable factor analysis
- Fast PARAFAC with imposed nonnegativity renders reasonable pure component factors for most species
  - Time shifting algorithm appears to make a suitable correction for follow-on analysis
  - Predicted pure component emission spectra and decay curves are in good agreement with single species solution data.
- PARAFAC2 is appropriate for time-shifted data, but inability to impose nonnegativity in time domain results in poor factor resolution
- Core PARAFAC2 with nonnegativity constraints provides results highly comparable to the shift-corrected data
  - Demonstrated excellent results in both the simulated and real data

Future Work

- Investigate methods to “oversample” core array
  - Using factor ranks greater than model rank
  - Easily done in Core PARAFAC
- Determine method to optimize number of PARAFAC cycles utilized on core
- Compare Core PARAFAC2 with other new methods

Acknowledgements

- The Sandia National Laboratories Laboratory Directed Research and Development (LDRD) program provided funding for this project.
- Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.
- Fluorescence data provided by Timothy Keller and Gregory D. Gillispie with financial support from the National Institutes of Health Small Business Innovations Research Program, NIHSBIR-1R43EB007866; and the Montana Board of Research and Commercialization Technology, grant 08-48