Local Functional Constraints in Multivariate Curve Resolution

Jeremy M. Shaver, Neal B. Gallagher
Eigenvector Research, Inc.
Manson, WA 98155
Rasmus Bro
Royal Veterinary and Agricultural University
Frederiksberg, Denmark

Shaver@Eigenvector.com

EIGENVECTOR RESEARCH INCORPORATED

Multivariate Curve Resolution

- MCR is very often used with spectra,
  - also known as self-modeling curve resolution, self-modeling mixture analysis, “end member extraction”
- Often used when you do NOT have calibration data
- Goal is to recover underlying "factors" which represent physically-identifiable features.

EIGENVECTOR RESEARCH INCORPORATED

2-Way MCR

- Based on the classical least squares (CLS) model, attempt to estimate **C** and **S** given **X**:

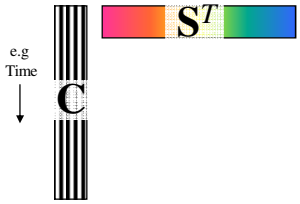
$$\mathbf{X} = \mathbf{C}\mathbf{S}^T + \mathbf{E}$$

X *M*x*N* measured responses,

C *M*x*K* pure analyte contributions,

S *N*x*K* pure analyte spectra, and

E *M*x*N* residuals.



- Non-Negative Alternating Least Squares (NN-ALS)

EIGENVECTOR RESEARCH INCORPORATED

Outline

- Purity Initialization Method (DISTSLCT)
- Raman Image Example
  - Sequential Initialization
- N-Way Purity
- Unfolding, selecting (DISTSLCTN)
- Predicting the non-selective mode
- Examples
- Conclusions

EIGENVECTOR RESEARCH INCORPORATED

Purity Method

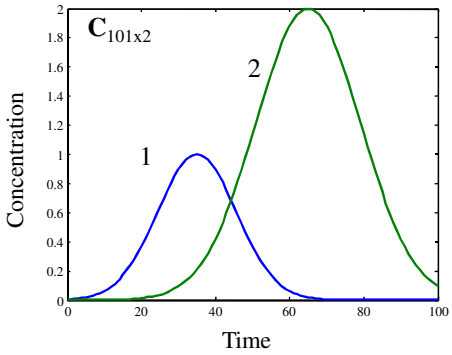
- Uses a geometrical argument
- Samples (variables) on the exterior of the data cloud are representative of "pure" responses
  - true if there is a selective sample (variable)

EIGENVECTOR RESEARCH INCORPORATED

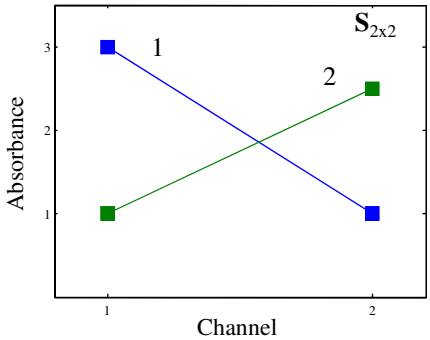
Example with Evolving Data

- Synthetic data from LC-NIR

Elution Profiles

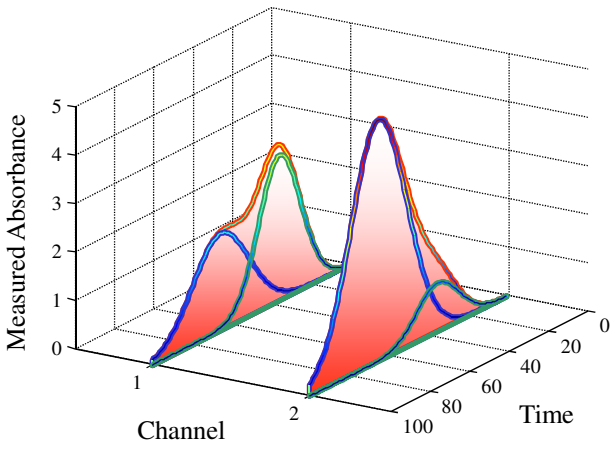


2-Channel spectra



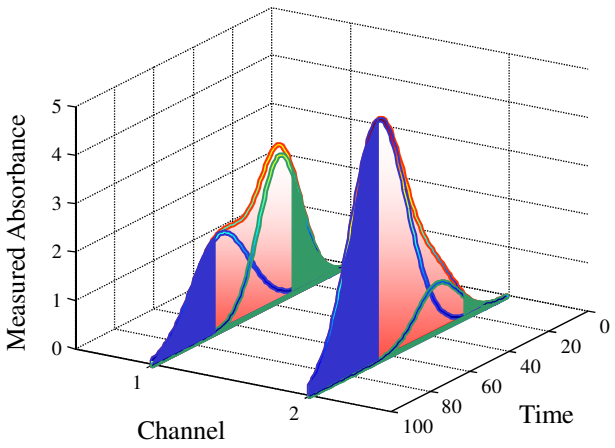
EIGENVECTOR RESEARCH INCORPORATED

Measured Response (no noise)



EIGENVECTOR RESEARCH INCORPORATED

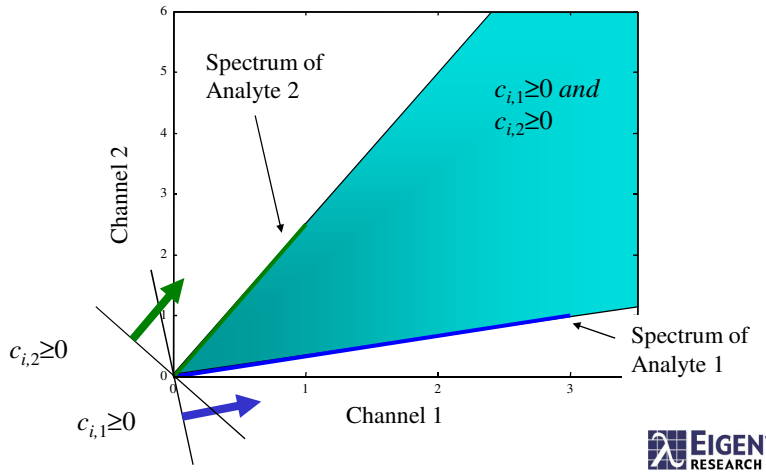
Measured Response (no noise)



EIGENVECTOR RESEARCH INCORPORATED

Non-negativity

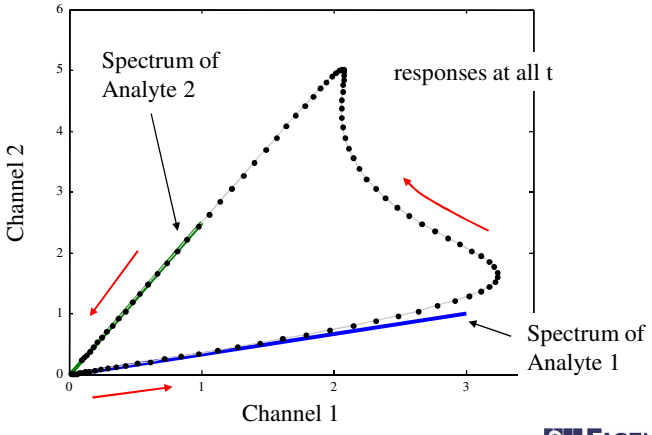
- Plot **S** (Channel 2 versus 1)



EIGENVECTOR RESEARCH INCORPORATED

Elution Example No Noise

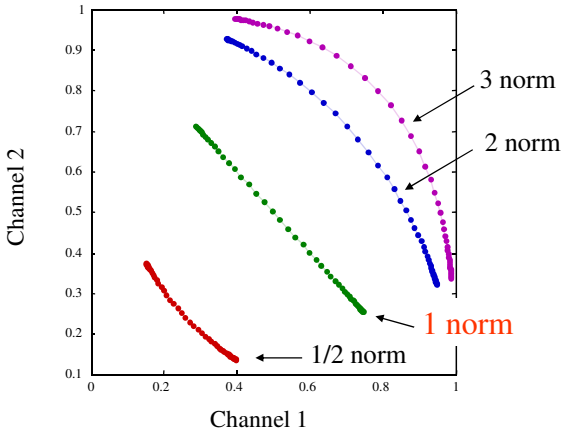
- Samples at the boundaries (extremes) are best estimate for $\mathbf{S}_0$



EIGENVECTOR RESEARCH INCORPORATED

How to find the Extremes?

- Normalize each spectrum (which p?)



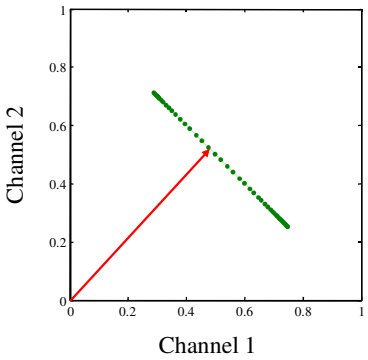
p norm

$$\mathbf{x} = \mathbf{x}\left(\sum_{j=1}^{N} x_j^p\right)^{1/p}$$

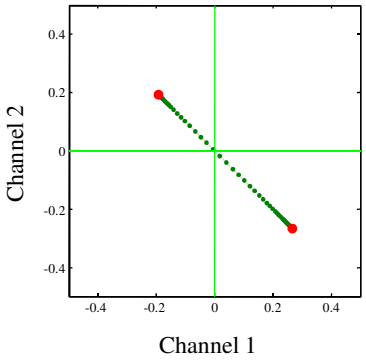
EIGENVECTOR RESEARCH INCORPORATED

Extreme Samples

- Mean centering the 1 norm spectra drops the rank



the extreme sample spectra are indistinguishable from the original analyte spectra

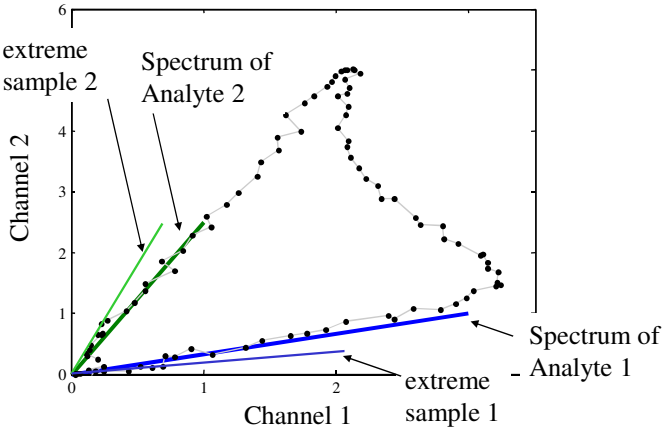


samples with 0 norm not used

EIGENVECTOR RESEARCH INCORPORATED

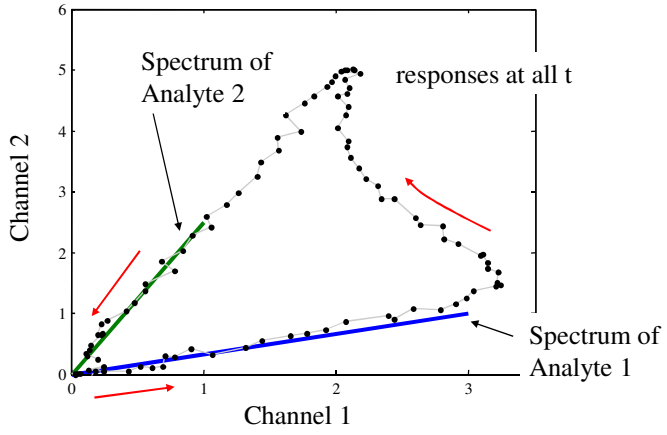
What Happens with Noise?

- Noise pushes the data cloud boundaries outward



What Happens with Noise?

- Estimate extremes using samples with higher signal (e.g. from samples with norm >0.5)



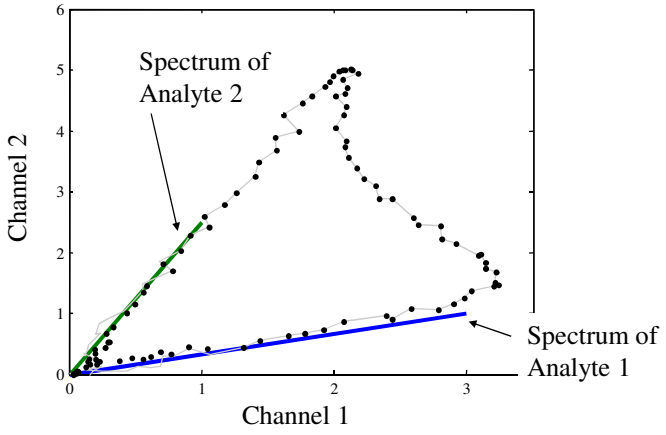
Offsets

- Adding a small offset to each spectrum prior to normalization moves *all* spectra towards the center of the data cloud
  - low signal (high noise) spectra are moved more than high signal samples
  - assumes that low signal spectra are noisiest
  - offset can be selected that is ~noise level
    - default ~1-3%

EIGENVECTOR RESEARCH INCORPORATED

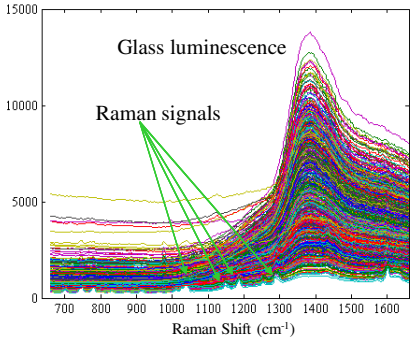
Offset Added

- Estimate boundaries using samples with higher signal



Raman Image

- Aspirin and polyethylene on a glass slide
- Raman 21*x*33 *x* 501
  660-1660 cm^{-1}
- Background
  luminescence varies for each pixel



Courtesy Kaiser Optical Systems

EIGENVECTOR RESEARCH INCORPORATED

Aspirin DISTSLCT Samples

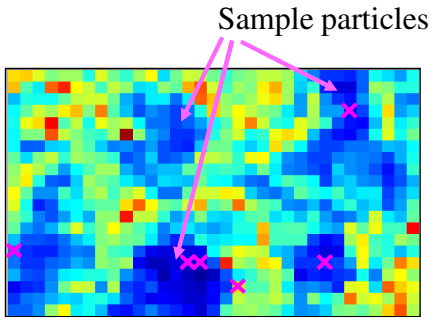
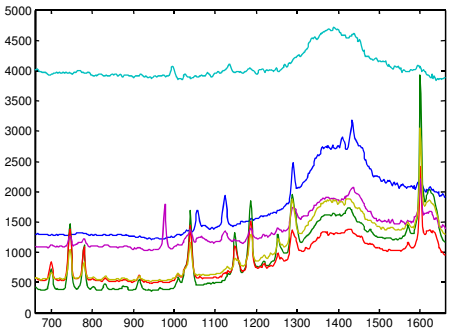


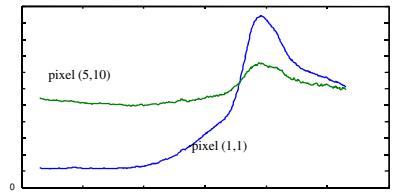
Image of summed intensity

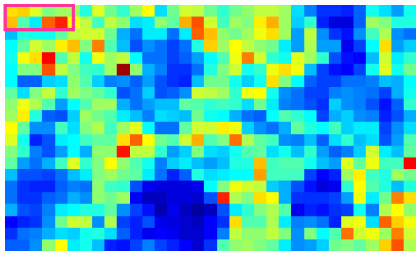


EIGENVECTOR RESEARCH INCORPORATED

After Initial Run...

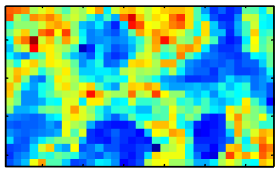
- Use unique Variables
- 6 Apparent Factors
- Add fixed offset component.
- Fix some background pixels to be single component.

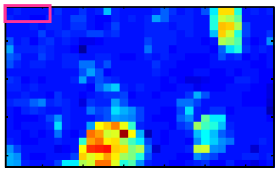


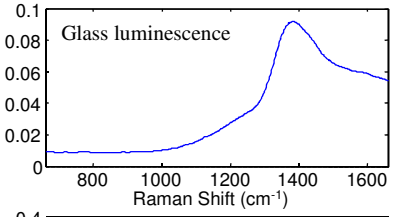


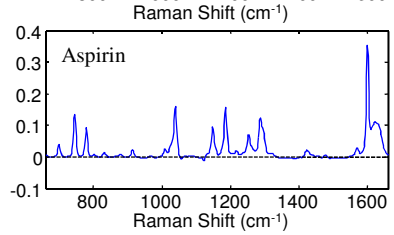
EIGENVECTOR RESEARCH INCORPORATED

Non-Negative MCR + Equality Constraints



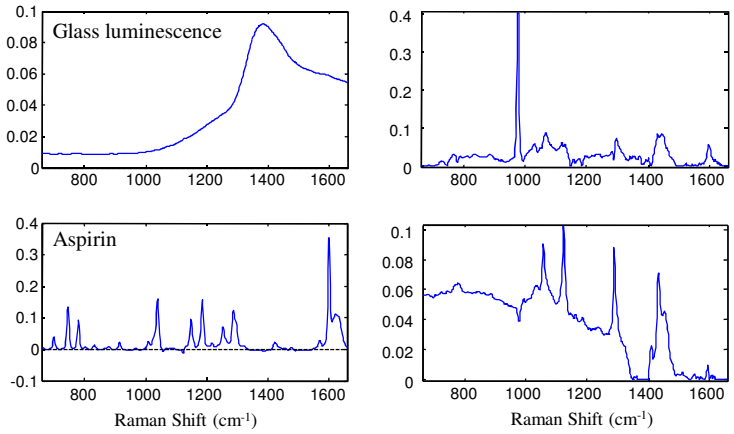






EIGENVECTOR RESEARCH INCORPORATED

Non-Negative MCR + Equality Constraints

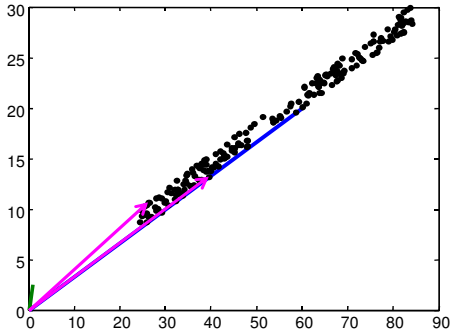


+ 1 high- and 1 low-signal background

EIGENVECTOR RESEARCH INCORPORATED

Distribution of Samples

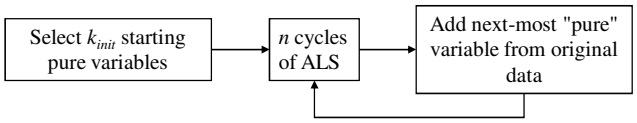
- **Problems:**
  - Samples do not effectively span space
  - Some components have exceptionally small proportional contributions
- **Solution: Hmmmmm…**



EIGENVECTOR RESEARCH INCORPORATED

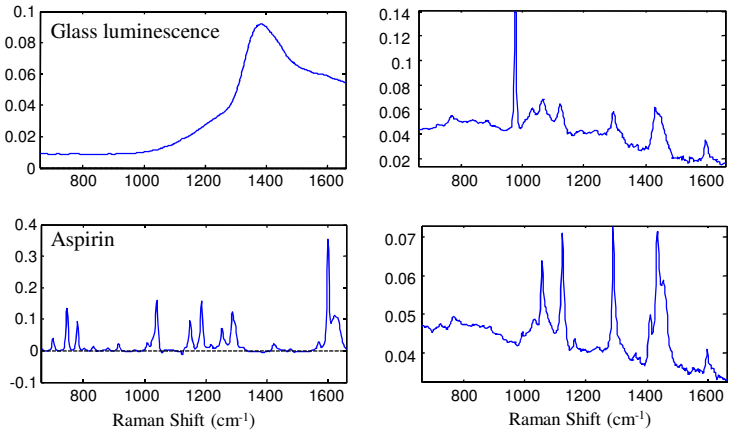
Handling Low Signal Components

- Use "Sequential" MCR
  - Limit to initial (high variance) components
  - Add additional components after several iterations
  - Proven very effective for exceptionally high-rank systems!



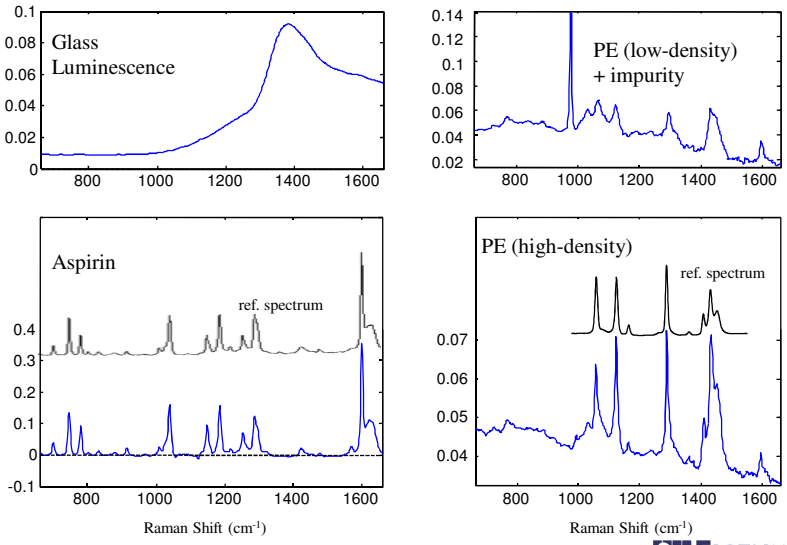
EIGENVECTOR RESEARCH INCORPORATED

Sequential MCR Results



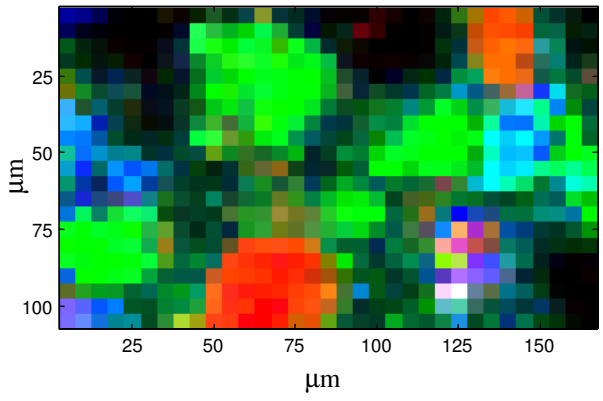
EIGENVECTOR RESEARCH INCORPORATED

Recovered Spectra



EIGENVECTOR RESEARCH INCORPORATED

Recovered Distributions

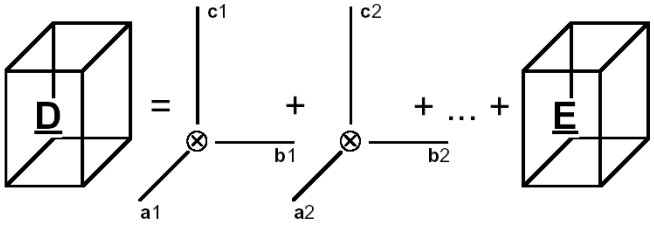


Aspirin
High-density polyethylene
Low-density polyethylene

EIGENVECTOR RESEARCH INCORPORATED

PARAFAC N-way Model

$$d_{ijk} = \Sigma_q a_{iq} b_{jq} c_{kq} + e_{ijk}$$



EIGENVECTOR RESEARCH INCORPORATED

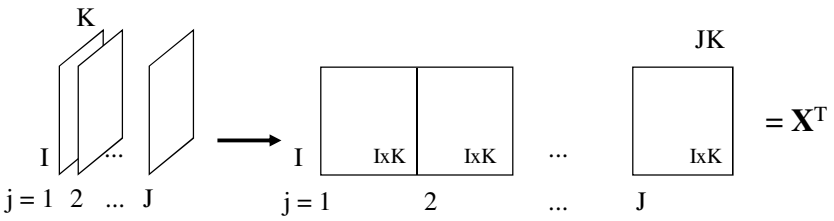
Extend to N-way

- In 2-way, extremes are selected (exterior of data cloud) for *either* Samples *or* Variables
  - predict the other mode using least squares
  - not least squares for first mode
- In N-way, extremes are selected for N-1 modes
  - then predict the mode left out using least squares
  - not least squares for N-1 modes
  - requires unfolding

EIGENVECTOR RESEARCH INCORPORATED

"Pure" Estimate for Mode 1

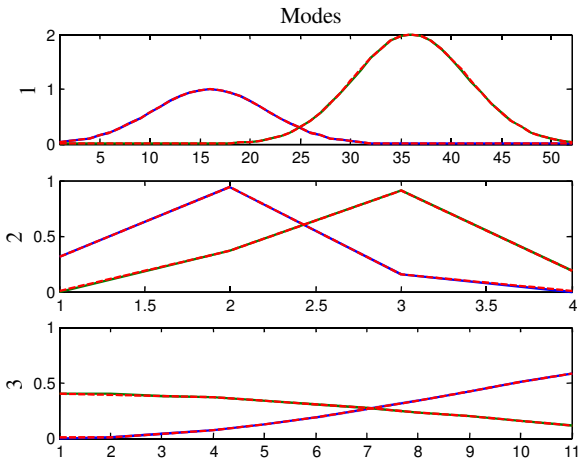
- Normalize the rows of **X** to give $\mathbf{X}_{norm}$
  - 1-norm
- Mean center the columns of $\mathbf{X}_{norm}$ to give $\mathbf{X}_{norm,center}$



EIGENVECTOR RESEARCH INCORPORATED

Synthetic Data

- 52x4x11
- Selective
- no-noise



EIGENVECTOR RESEARCH INCORPORATED

Amino Acids

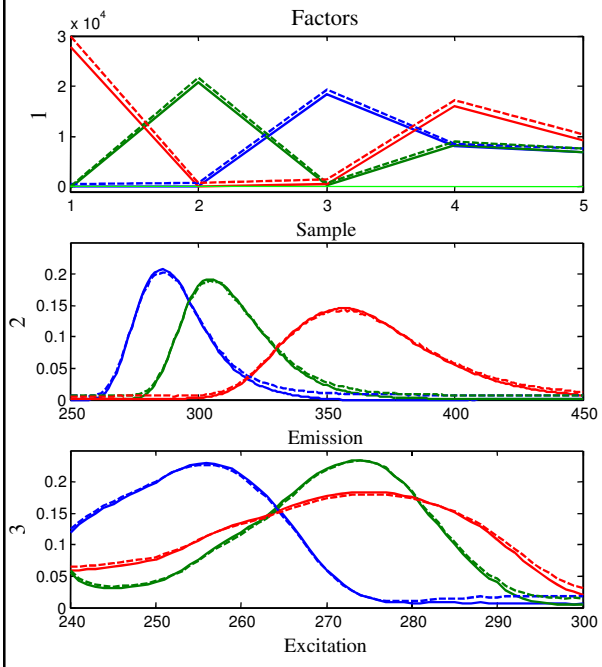
- Excitation/Emission

  Measured by Claus A. Andersson, described in Bro, R., Multi-way Analysis in the Food Industry. Models, Algorithms, and Applications. 1998. Ph.D. Thesis, University of Amsterdam (NL) and Royal Veterinary and Agricultural University (DK)

  In PLS_Toolbox Version 3

- 5 x 201 x 61
  - 5 samples
  - 291 emission wavelengths
  - 61 excitation wavelengths (non-selective mode)
  - offset = 1%

EIGENVECTOR RESEARCH INCORPORATED



Amino Acids

PARAFAC

DISTSLCTN

EIGENVECTOR RESEARCH INCORPORATED

Sugar Data

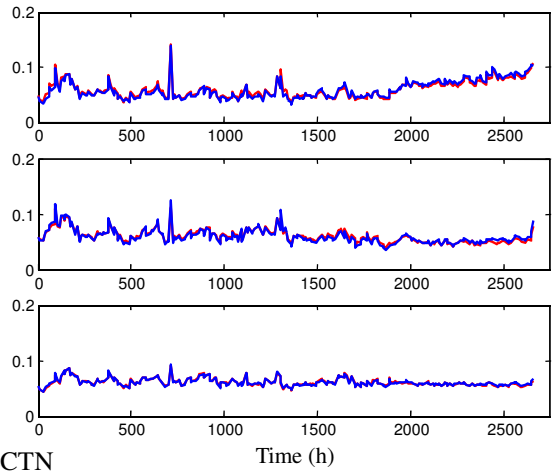
- Excitation/Emission
- Slightly different version of

  R. Bro. Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis. *Chemom. Intell. Lab. Syst.* 46:133-147, 1999.

  in PLS_Toolbox Version 3
- 268 x 44 x 7
  - 268 sample times
  - 44 emission wavelengths
  - 7 excitation wavelengths (non-selective mode)

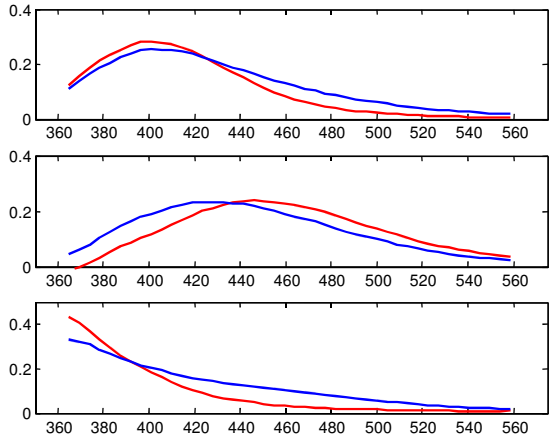
EIGENVECTOR RESEARCH INCORPORATED

Sugar Data: Mode 1



EIGENVECTOR RESEARCH INCORPORATED

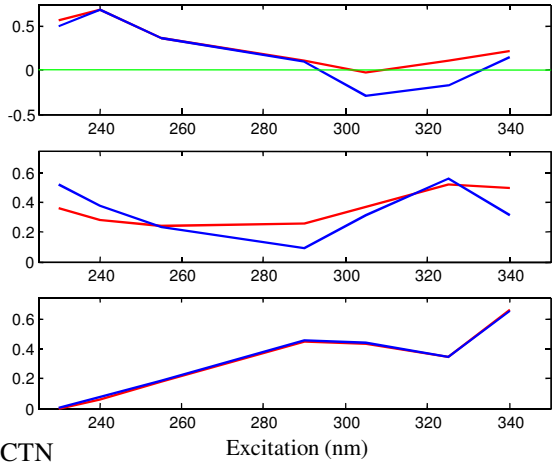
Sugar Data: Mode 2



DISTSLCTN
PARAFAC

EIGENVECTOR RESEARCH INCORPORATED

Sugar Data: Mode 3



DISTSLCTN
PARAFAC

EIGENVECTOR RESEARCH INCORPORATED

Time Comparison

- PARAFAC
  - ALS
  - Init'zed w/ TLD
  - error trapping, more consistency checking overhead, display
- DISTSLCTN
  - Purity
  - less overhead, less error trapping, fewer options and no constraints
  - offset 1% of max

EIGENVECTOR RESEARCH INCORPORATED

Compare TLD and DISTSLCTN

	Time (s)*		
i	DIST	TLD	PARAFAC/(iterations)
Amino	0.5	1.3	5.8 /(35)
Sugar	0.6	1.9	59 /(250)

** HP Vectra , P4-1.7 GHz, 1 Gb RAM, Win2K
PLS_Toolbox, Ver 3.0.2
MATLAB, Ver 6.5

EIGENVECTOR RESEARCH INCORPORATED

Angle Between Factors

- For factor i

$$\mathbf{f}_i = vec\left(\mathbf{a}_i \otimes \mathbf{b}_i \otimes \mathbf{c}_i\right)$$

- The angle between factor i for method A and PARAFAC factors P

$$\theta_{i,AP} = \cos^{-1}\left(\frac{\mathbf{f}_{i,A}^T}{\left|\mathbf{f}_{i,A}\right|}\frac{\mathbf{f}_{i,P}}{\left|\mathbf{f}_{i,P}\right|}\right)$$

EIGENVECTOR RESEARCH INCORPORATED

Compare TLD and DISTSLCTN

Angle with PARAFAC Factors, $\theta_{i,AP}$

Factor	Amino		Sugar	
i	TLD	DIST	TLD	DIST
1	10	3	7	33
2	9	3	7	29
3	1	3	4	21

EIGENVECTOR RESEARCH INCORPORATED

Conclusions

- A sequential version of purity-based initialization for MCR has shown promise in certain situations
- A Purity-based method has been used to extract factors for N-way ("Pure" PARAFAC)
  - quick initialization (can examine many models)
- DISTSLCTN is
  - slightly faster than TLD
  - good first estimates
    - but not as good as TLD when selectivity poor
  - can be applied to N-way
    - TLD is for 3-way

EIGENVECTOR RESEARCH INCORPORATED

Future Work

- Factor Matching
  - In purity, the selective modes are independent which can result in "un-matched" factors
  - Present method matches to first mode using a minimization of residuals
- Missing data
  - E.g. EEM
- Allow for interactive selection

EIGENVECTOR RESEARCH INCORPORATED

Where Does It Lead?

- Many people developing ALS modifications or data pre-treatments
  - R. Tauler, A. Smilde, and B. Kowalski, J. Chemometrics, **9**, 31-58 (1995)
  - MH Van Benthem, MR Keenan, DM Haaland, *J. Chemometrics*, **16**, 613-622 (2002)
  - P.J. Gemperline, E. Cash, *Anal. Chem.*, **75**, 4236-4243 (2003)
- Better characterization of uncertainty
  - R. Tauler, *J. Chemometrics*, **15**, 627-646 (2001)
- Other methods also used with success
  - W. Windig, et. al., *Anal. Chem* **74**, 1371-1379 (2002)

EIGENVECTOR RESEARCH INCORPORATED

Where Does It Lead?

- "Technique Overload" Question
  - So many to choose from!?!
  - Often evaluated from an empirical point of view…Why did it work?
  - How to compare results?
- "Silver Bullet" Question
  - Can these techniques be combined into a single amalgamated method that will work on *almost all* problems?
  - How to automate integration?

EIGENVECTOR RESEARCH INCORPORATED