Purity-Based Method for Initializing PARAFAC

N.B. Gallagher
Eigenvector Research, Inc.
nealg@eigenvector.com

EIGENVECTOR RESEARCH INCORPORATED

Outline

- Purity method
- Unfolding, selecting (DISTSLCTN)
- Predicting the non-selective mode
- Examples
- Conclusions

EIGENVECTOR RESEARCH INCORPORATED

Purity Method

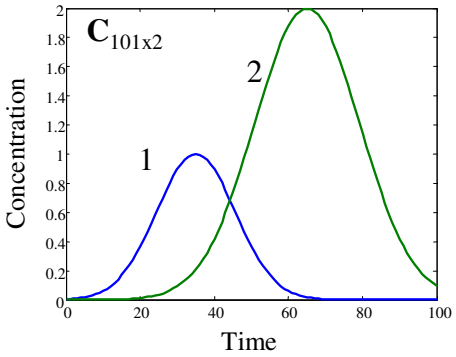
- Uses a geometrical argument
- Samples (variables) on the exterior of the data cloud are representative of "pure" responses
  - true if there is a selective sample (variable)

EIGENVECTOR RESEARCH INCORPORATED

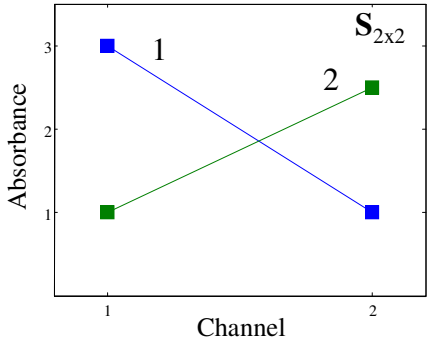
Example with Evolving Data

- Synthetic data from LC-NIR

Elution Profiles

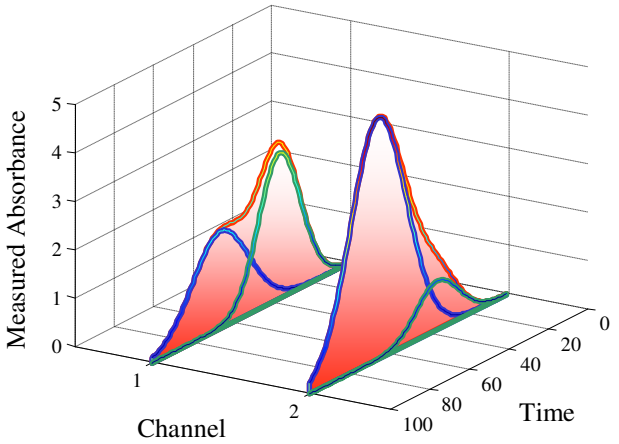


2-Channel spectra



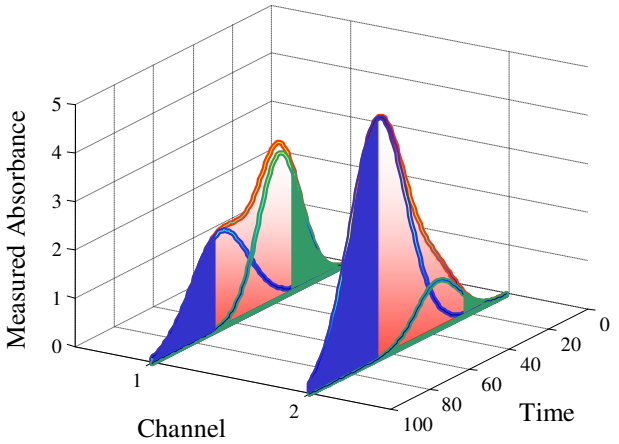
EIGENVECTOR RESEARCH INCORPORATED

Measured Response (no noise)



EIGENVECTOR RESEARCH INCORPORATED

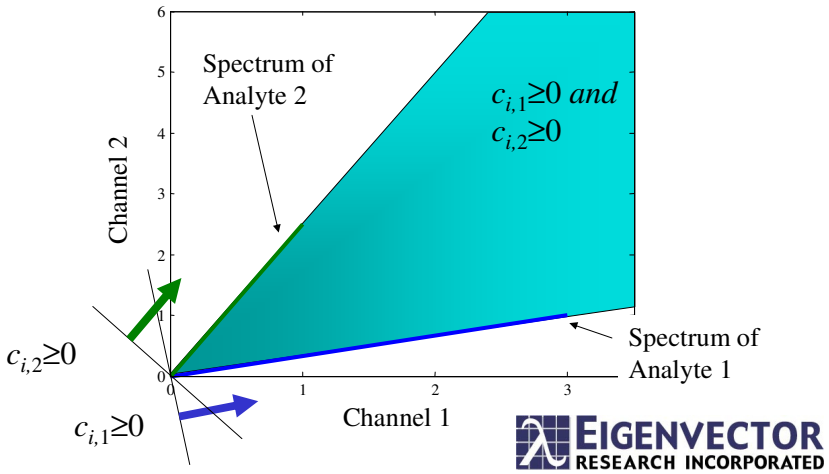
Measured Response (no noise)



EIGENVECTOR RESEARCH INCORPORATED

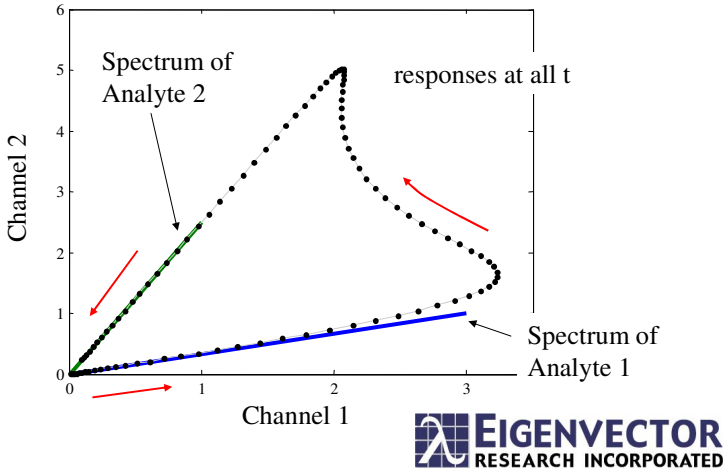
Non-negativity

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



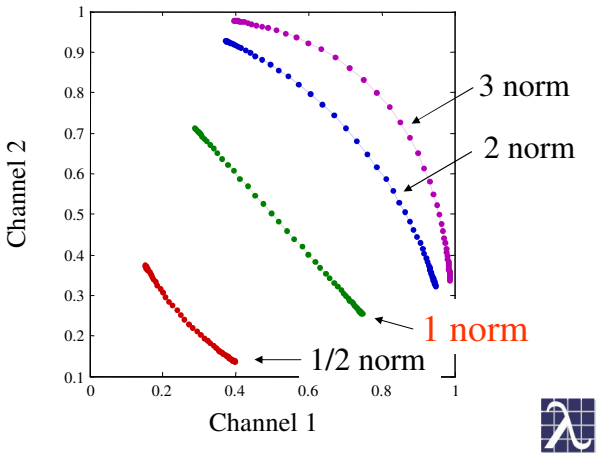
Elution Example No Noise

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



How to find the Extremes?

- Normalize each spectrum (which *p*?)



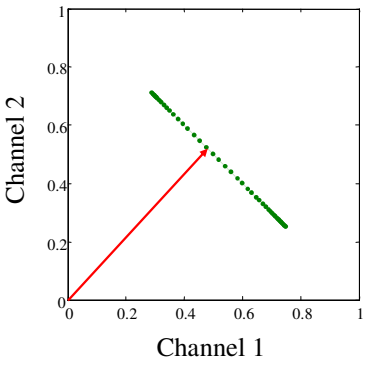
p norm

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

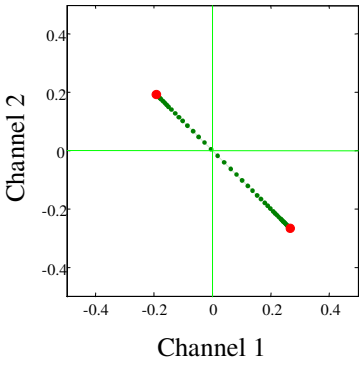
EIGENVECTOR RESEARCH INCORPORATED

Extreme Samples

- Mean centering the 1 norm spectra drops the rank



the extreme sample spectra are indistinguishable from the original analyte spectra

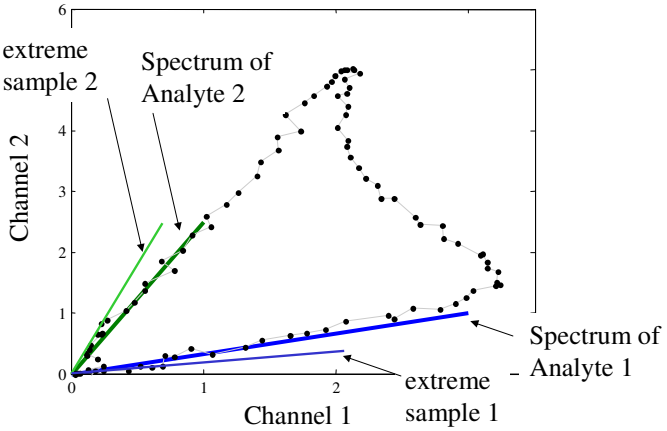


samples with 0 norm not used

EIGENVECTOR RESEARCH INCORPORATED

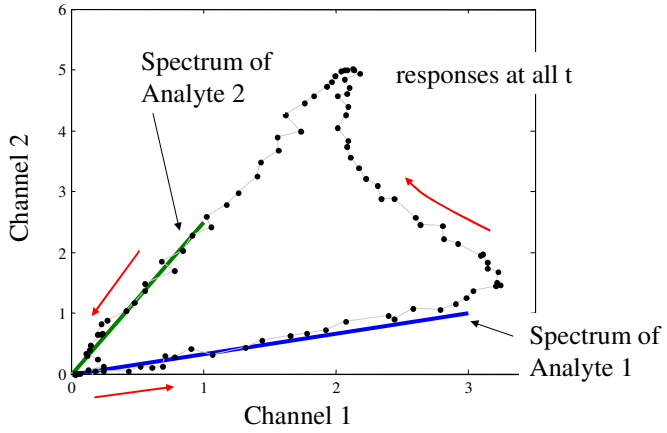
What Happens with Noise?

- 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)



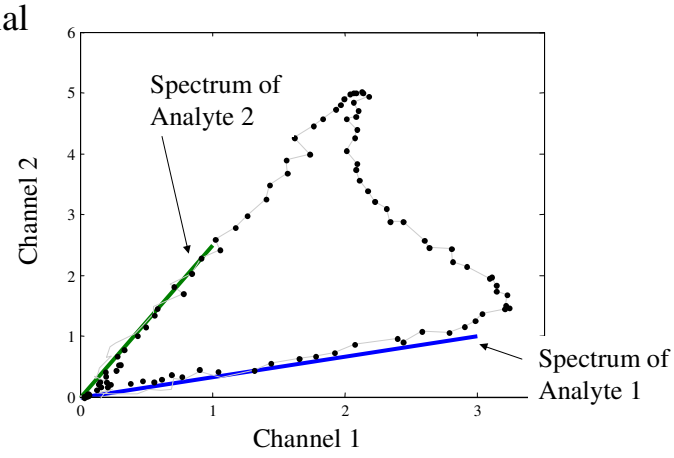
Off-sets

- Adding a small off-set 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
  - off-set can be selected that is ~noise level
    - default ~1-3%

EIGENVECTOR RESEARCH INCORPORATED

Off-set Added

- Estimate boundaries using samples with higher signal



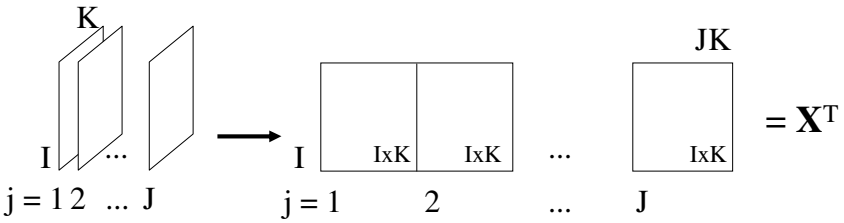
Extend to N-way

- In 2-way, extremes are selected (exterior of data cloud) for *either* Rows *or* Variables
  - then 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

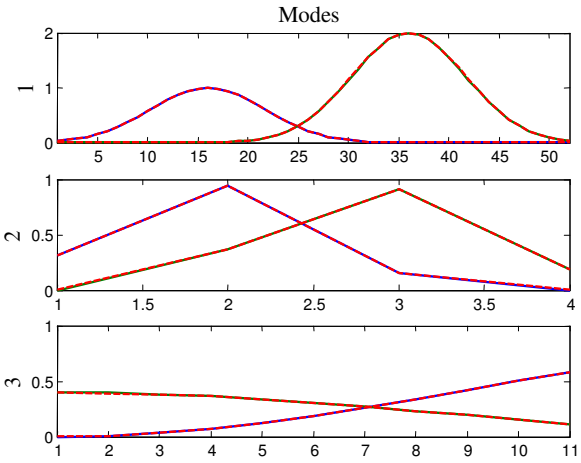
- Normalize the rows of **X** to give $\mathbf{X}_{norm}$
  - 1-norm
- Mean center the columns of $\mathbf{X}_{norm}$ to give $\mathbf{X}_{norm,center}$



EIGENVECTOR RESEARCH INCORPORATED

Synthetic Data

- 52x4x11
- Selective
- no-noise



EIGENVECTOR RESEARCH INCORPORATED

Amino Acids

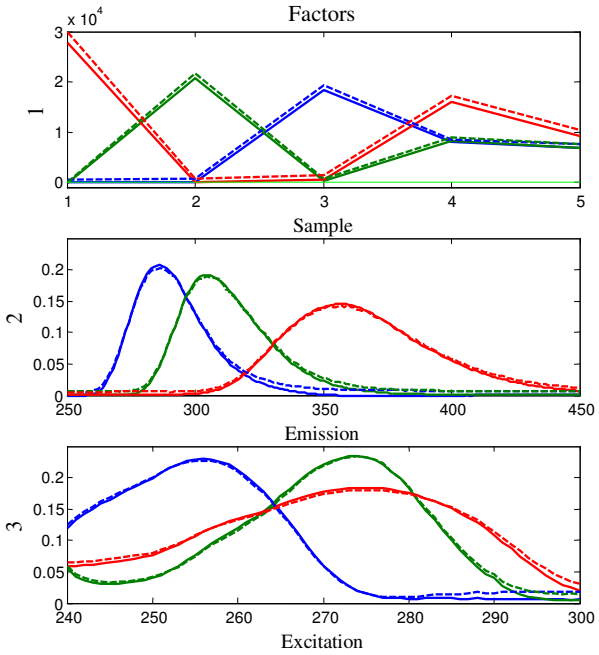
- Excitation/emission

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

solid: PARAFAC

dashed: DISTSLCTN

EIGENVECTOR RESEARCH INCORPORATED

Sugar Data

- Excitation/emission
- Slightly different version of

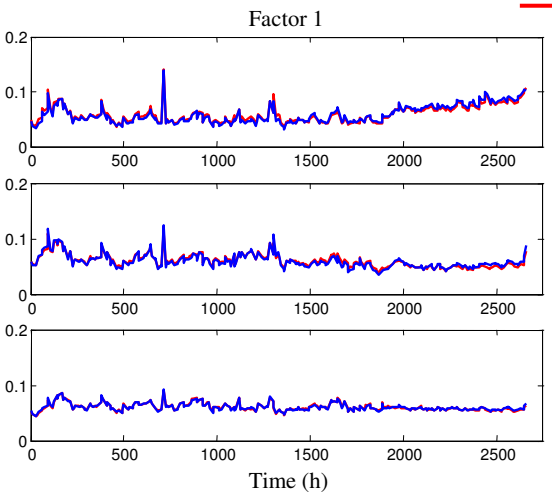
  R. Bro. Exploratory study of sugar production using fluorescence spectroscopy and multi-way analysis. *Chemom. Intell. Lab. Syst.* 46:133-147, 1999.

  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: Factor 1

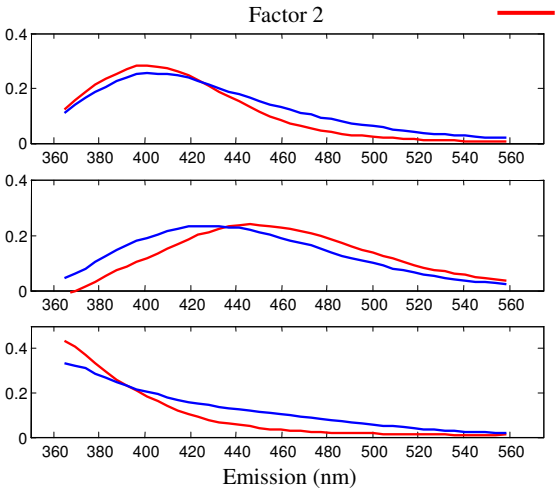
DISTSLCTN
PARAFAC



EIGENVECTOR RESEARCH INCORPORATED

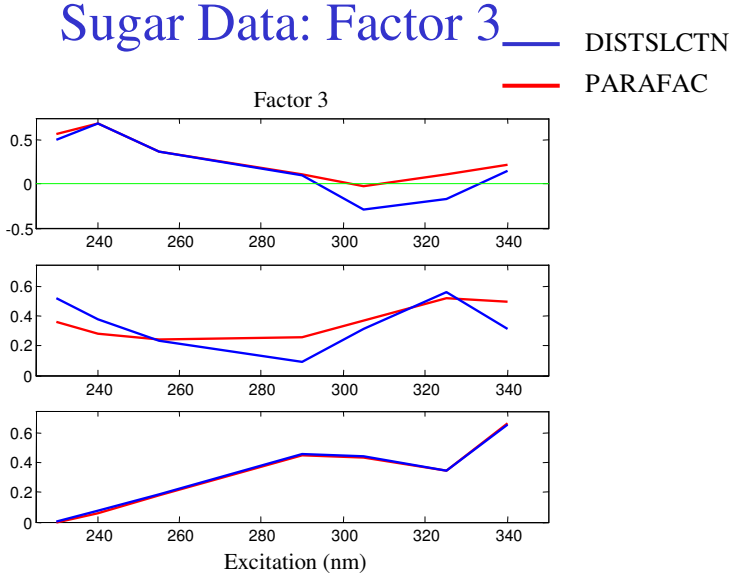
Sugar Data: Factor 2

DISTSLCTN
PARAFAC



EIGENVECTOR RESEARCH INCORPORATED

Sugar Data: Factor 3



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 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

EIGENVECTOR RESEARCH INCORPORATED

Fluor Data

- Excitation/emission: 4-way
- 6 Factors (Catechol, Hydroquinone, Indole, Resorcinol, Tryptophane, Tyrosine)
  - Reyleigh scatter replaced with PCA for each sample

  Data acquired by Åsmund Rinnan. KVL
- 358 x 136 x 19 x 5
  - 358 samples (only non-zero conc used)
  - 136 emission wavelengths
  - 19 excitation wavelengths
  - 5 replicates (non-selective mode)

EIGENVECTOR RESEARCH INCORPORATED

Mode 1 and 4

- Sample and Replicate Mode

EIGENVECTOR RESEARCH INCORPORATED