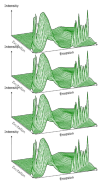
Introduction to Multiway Analysis

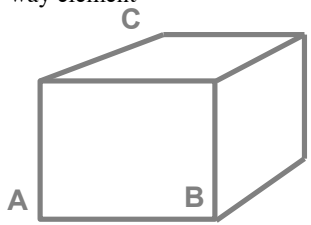
EIGENVECTOR RESEARCH INCORPORATED

Introduction

Three-way data?

- Simply a set of 'equivalent' two-way matrices obtained at different occasions
- Data measured as a function of three 'things'
  - E.g. samples, variables, times
- x_{ij} is a matrix element and x_{ijk} is a three-way element





EIGENVECTOR RESEARCH INCORPORATED

Examples

Sensory analysis
- Score as a function of (Food sample, Judge, Attribute)

Process analysis
- Measurement as a function of (Batch, Variable, time)
- Measurement as a function of (Variable, Lag, Location)

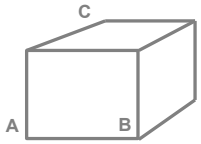
Image analysis
- Pixelvalue as a function of (Sample, Image pixel, Variable)

Experimental design
- Response as a function of (factor 1, factor2, factor3,..)

Spectroscopy
- Intensity as a function of (Wavelength, Retention, Sample, Time, Location , Treatment)

Environmental analysis
- Measurement as a function of (Location, Time, Variable)

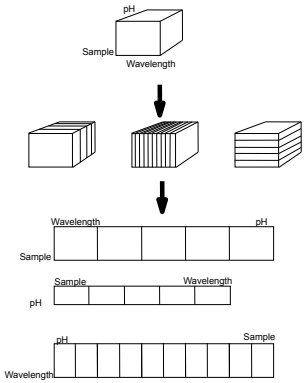
Chromatography
- Measurement as a function of (Sample, Retention time, Variable)



Unfolding/matricization

Traditional approach
- Unfolding leading to two-way data and analysis

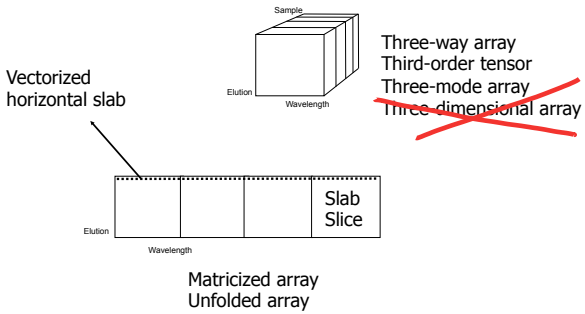
Three-way models
- Natural extensions of two-way models
- PCA leads to PARAFAC or Tucker3 depending on how it is extended
- PLS leads to multilinear PLS (N-PLS)



Unfolding/matriciation
Often leads to overfitting

EIGENVECTOR RESEARCH INCORPORATED

Multi-way notation



Kiers. Towards a standardized notation and terminology in multiway analysis. *Journal of Chemometrics* 14 (3):105-122, 2000.

EIGENVECTOR RESEARCH INCORPORATED

PARAFAC

- PCA - bilinear model,

$$x_{ij} = \sum_{f=1}^{F} a_{if} b_{jf} + e_{ij}$$

EIGENVECTOR RESEARCH INCORPORATED

PARAFAC

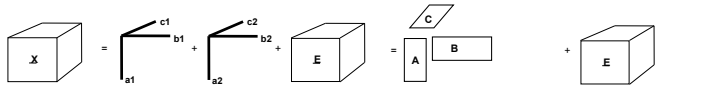
PARAFAC invented in 1970 by Harshman based on Cattell 1944

- PCA - bilinear model,

$$x_{ij} = \sum_{f=1}^{F} a_{if} b_{jf} + e_{ij}$$

- PARAFAC - trilinear model,

$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk}$$



EIGENVECTOR RESEARCH INCORPORATED

PARAFAC

…. or

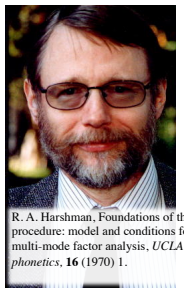
Canonical Decomposition

=

CanDecomp

…or combined

CP

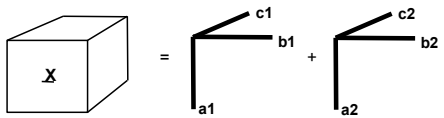


R. A. Harshman, Foundations of the PARAFAC procedure: model and conditions for an 'explanatory' multi-mode factor analysis, *UCLA Working Papers in phonetics*, **16** (1970) 1.

J. D. Carroll, J. Chang, Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition, *Psychometrika*, **35** (1970) 283.

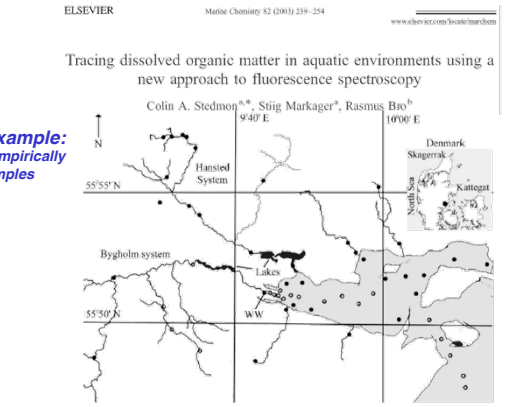
Three-way rank

- Rank of two-way matrix
  - Minimum number of bilinear (PCA) components needed to reproduce matrix
- Rank of three-way array
  - Minimum number of trilinear (PARAFAC) components needed to reproduce array



EIGENVECTOR RESEARCH INCORPORATED

A motivating example:
Monitor pollution empirically from water samples


ELSEVIER

Marine Chemistry 82 (2003) 239–254

www.elsevier.com/locate/marchem

Tracing dissolved organic matter in aquatic environments using a new approach to fluorescence spectroscopy

Colin A. Stedmon[a,*], Stiig Markager[a], Rasmus Bro[b]

9°40' E

10°00' E

N

55°55' N

Hansted System

Denmark

Skagerrak

North Sea

Kattegat

Bygholm system

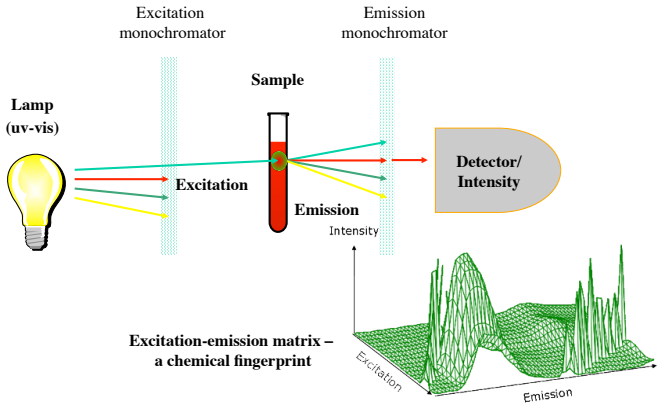
Lakes

55°50' N

WW

Fig. 1. Map of the Horsens catchment and its location on the east coast of Jutland, Denmark. Filled circles represent sampling stations and hollow circles represent the locations of the samples used in Fig. 7.

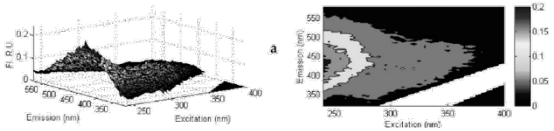
EIGENVECTOR RESEARCH INCORPORATED



Fluorescence

EIGENVECTOR RESEARCH INCORPORATED

A practical example: Data from each sample

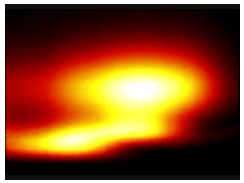


State of the art: Peak-spotting!!!

EIGENVECTOR RESEARCH INCORPORATED

Practical example

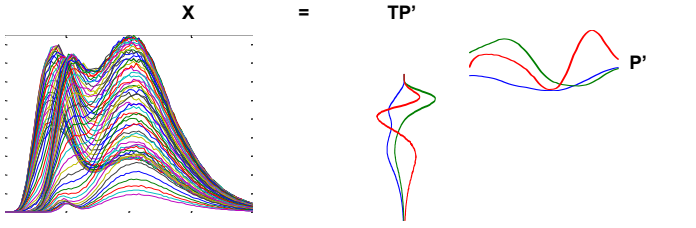
- Rotational freedom of PCA



EIGENVECTOR RESEARCH INCORPORATED

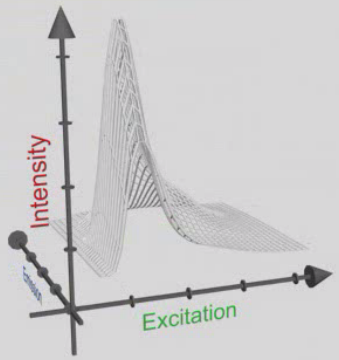
Practical example

- Rotational freedom of PCA



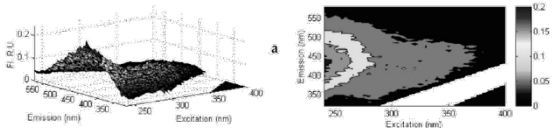
EIGENVECTOR RESEARCH INCORPORATED

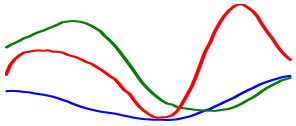
Food Technology - LMT - KVL - http://models.kvl.dk



Food Technology - LMT - KVL - http://models.kvl.dk

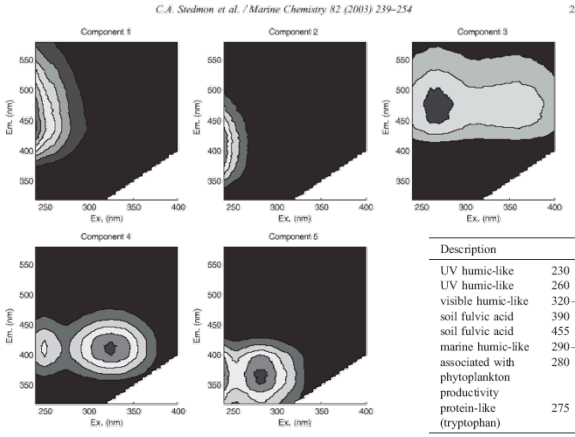
Instead of peak spotting or strange PCA components





EIGENVECTOR RESEARCH INCORPORATED

Get the real thing – the chemistry



Description		
UV humic-like	230	430
UV humic-like	260	380–460
visible humic-like	320–360	420–460
soil fulvic acid	390	509
soil fulvic acid	455	521
marine humic-like	290–310	370–410
associated with phytoplankton productivity	280	370
protein-like (tryptophan)	275	340

OR

EIGENVECTOR RESEARCH INCORPORATED

Mathematical chromatography

Uniqueness - what does it mean?

- Mixtures of analytes can be separated
- Concentrations can be estimated
- Pure spectra and profiles can be estimated

Eliminates major problems

- Removes indirect correlations
- Eliminates outliers
- Determines underlying sources
- Chemical/physical = simpler
- Way more noise insensitive

EIGENVECTOR RESEARCH INCORPORATED

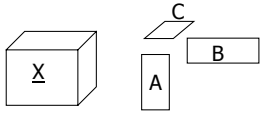
PARAFAC – when is it unique

Uniqueness* - conditions

A PARAFAC model is unique when

$$k_A + k_B + k_C \geq 2F + 2$$

F is the number of components and k_A is the k-rank of loading $\mathbf{A}$ = maximal number of randomly chosen columns which will have full rank ($\leq F$)



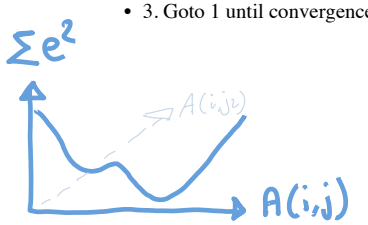
J. B. Kruskal. *Linear Algebra and its Applications* 18:95-138, 1977.
N. D. Sidiropoulos and R. Bro. *Journal of Chemometrics* 14 (3):229-239, 2000.

*Up to scaling and permutation

EIGENVECTOR RESEARCH INCORPORATED

- **PARAFAC algorithms *not* sequential**
  - PCA is a least squares model, fitted sequentially (NIPALS). Three comp. solution = two comp. plus one
  - PARAFAC *not* sequential. Hence refitting necessary.

- **Algorithm - Alternating least squares (ALS)**
  - Ex.: Bilinear model : $\|\mathbf{X} - \mathbf{AB}^T\|$
  - 1. $\mathbf{B}^T = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{X} = \mathbf{A}^+\mathbf{X}$
  - 2. $\mathbf{A}^T = (\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{X}^T = \mathbf{B}^+\mathbf{X}^T$
  - 3. Goto 1 until convergence (small change in fit $\|\mathbf{X}-\mathbf{AB}^T\|$)



PARAFAC Algorithm

1. Initialize **B** and **C**
2. $\mathbf{A} = \left(\sum_{k=1}^{K} \mathbf{X}_k\mathbf{B}\mathbf{D}_k\right)\left\{(\mathbf{B}'\mathbf{B}) * (\mathbf{C}'\mathbf{C})\right\}^{-1}$
3. $\mathbf{B} = \left(\sum_{k=1}^{K} \mathbf{X}'_k\mathbf{A}\mathbf{D}_k\right)\left\{(\mathbf{A}'\mathbf{A}) * (\mathbf{C}'\mathbf{C})\right\}^{-1}$
4. $diag\mathbf{D}_k = \left\{(\mathbf{B}'\mathbf{B}) * (\mathbf{A}'\mathbf{A})\right\}^{-1} \text{diag}(\mathbf{A}'\mathbf{X}_k\mathbf{B}), k=1,..,K$
5. Step 2 until relative change in fit is small

* Hadamard (elementwise product)

Why ALS?
Simple
Extends to N-way
Handles missing
Handles ML fitting
Constraints:
- Nonnegativity
- Unimodality
- Orthogonality
- Linear constraints
- Fixed parameters
- Smoothness
- Functional
- etc

Three-way ALS

EIGENVECTOR RESEARCH INCORPORATED

Two fundamental problems with PARAFAC

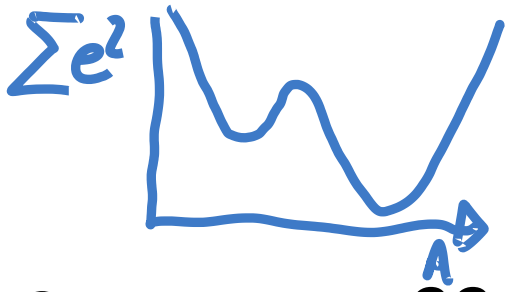
- Convergence
  - The solution may not be achieved because of lack of convergence
- Two-factor degeneracy
  - There may not be a solution at all

None of these problems occur in e.g. PCA.
When you do PCA, you 'get PCA'.

EIGENVECTOR RESEARCH INCORPORATED

The algorithm is iterative and stops when 'nothing happens'

Important to understand what that means and how it is evaluated

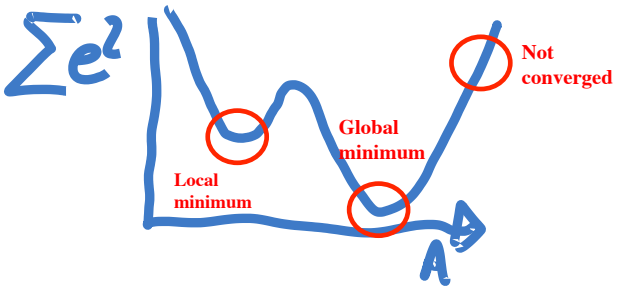


Convergence??

EIGENVECTOR RESEARCH INCORPORATED

The algorithm is iterative and stops when ‘nothing happens’

Important to understand what that means and how it is evaluated

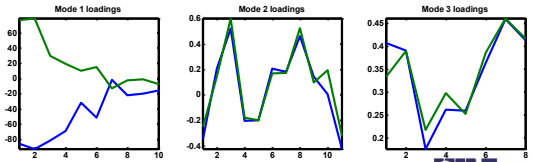


Convergence??

EIGENVECTOR RESEARCH INCORPORATED

Two-factor degeneracy

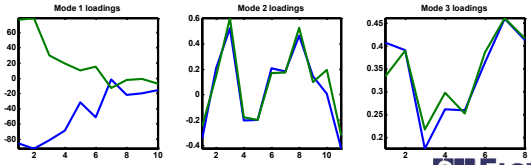
- Two factors become almost identical but with opposite sign ($a_1*b_1*c_1 = -a_2*b_2*c_2$)
- Grow in size and similarity with more iterations
- Combined contribution to the model is appr. zero
- Use Tuckers congruence (”correlation”) or plots to spot degeneracy



EIGENVECTOR RESEARCH INCORPORATED

Two-factor degeneracy

- Two factors become almost identical but with opposite sign ($a_1*b_1*c_1 = -a_2*b_2*c_2$)
- Happens when
  - Too many components or
  - When PARAFAC is not appropriate
  - Or during iterations something similar can happen



EIGENVECTOR RESEARCH INCORPORATED

Number of components?

Simpler than PCA (but takes more time):

- Cross-val, Scree etc. as in PCA
- Plus split-half
- Plus core consistency
- Plus chemical validation
- Plus algorithmic indications (degeneracy, many iterations, local minima etc.)

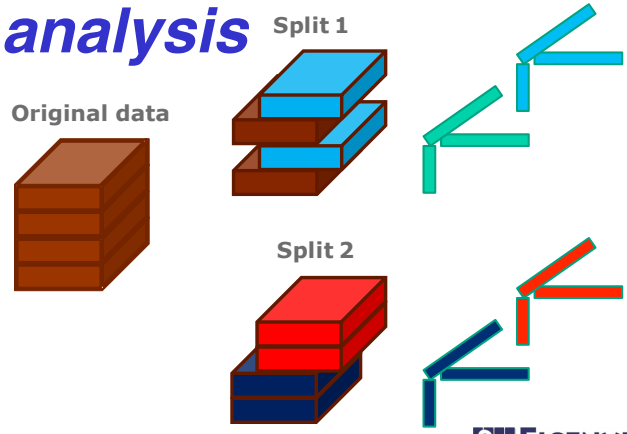
These are the main ones.
Always *look* at the model to validate it.
Use core consistency but carefully.
Use split-half for definitive validation

EIGENVECTOR RESEARCH INCORPORATED

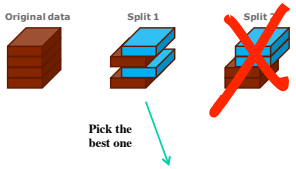
Core consistency

Split-half analysis

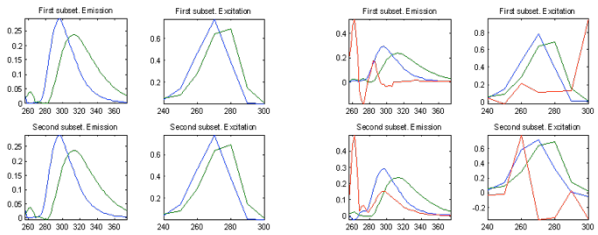
- Fit model to several independent samples
- If loadings the same, reasonable number chosen



Pick the best of the two



Split-half analysis



Multi-way analysis in fluorescence data, Report 2002, Giorgio Saracelle & Rasmus Bro

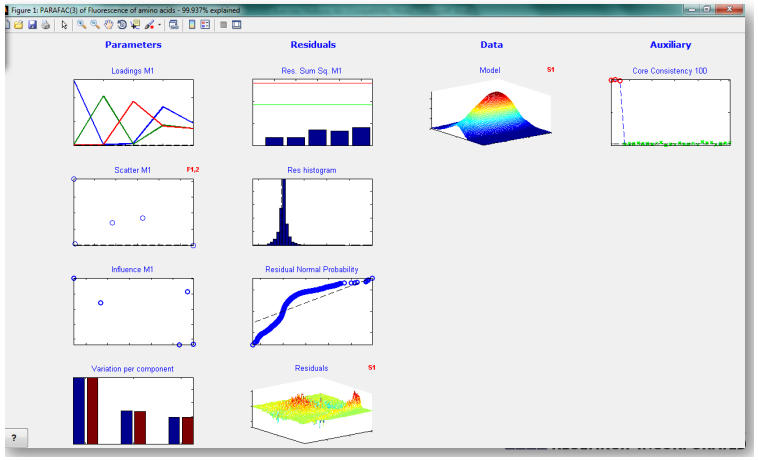
Same as in two-way analysis

- Residual analysis - $\mathbf{X} = \mathbf{M} + \mathbf{E}$, all of same sizes
  Look at (summed) squared residuals to find unusually large residuals

- Influence analysis - **A**, **B**, **C** from Tucker or PARAFAC Plot e.g. for finding extreme samples or use for calculating leverages (Hotellings T^2).

Outlier detection

Plots available



Some applications

EIGENVECTOR RESEARCH INCORPORATED

Second order calibration

How many samples are needed to build a calibration model?

Say you want to predict protein in flour. How many samples would you need?

EIGENVECTOR RESEARCH INCORPORATED

Curve resolution

For two-way data lots of ‘tricks’ necessary to fix rotational freedom

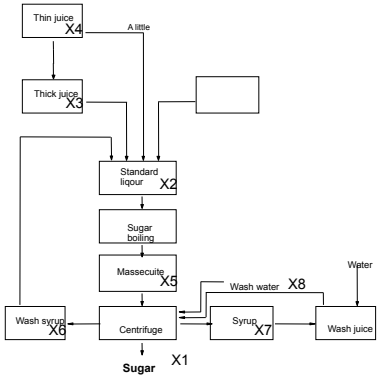
For three-way trilinear data …… PARAFAC

EIGENVECTOR RESEARCH INCORPORATED

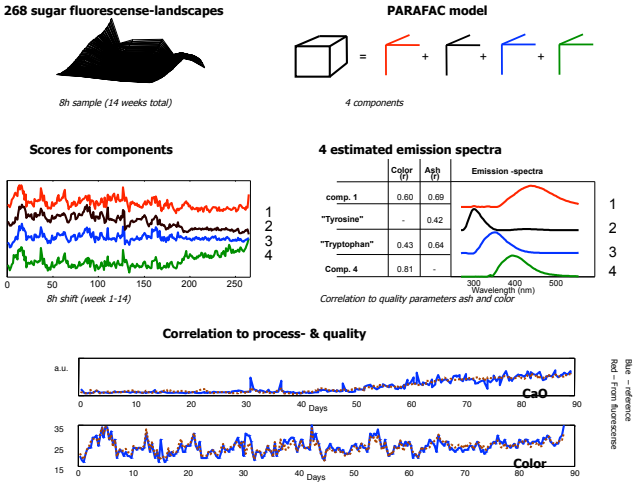
Sugar processing

Sugar made from beets

- Product sampled 8th hour for three months
- Fluorescence measured
- 260 samples



EIGENVECTOR RESEARCH INCORPORATED

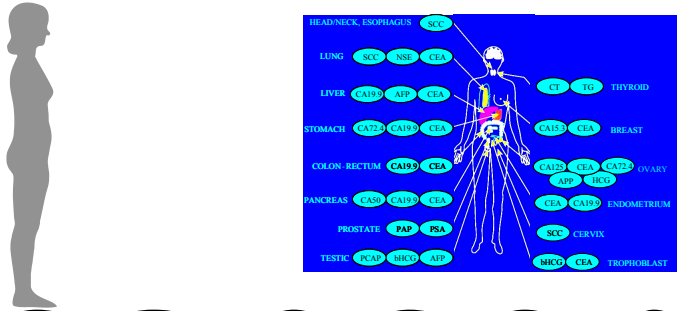


PARAFAC + fluorescence

- **Several advantages**
  - Chromatographic analysis of the whole process
  - Process monitoring (MSPC) on a chemical basis
  - Chemical understanding of why coloring occurs (PAC)
  - On-line prediction of quality
  - On-line prediction of process parameters
  - Nowadays = PAT – process analytical technology

EIGENVECTOR RESEARCH INCORPORATED

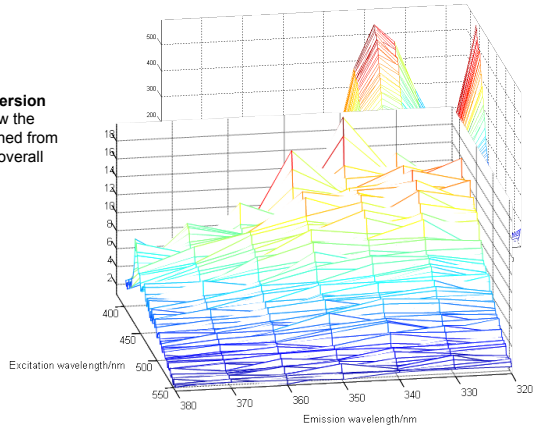
Traditional approach for cancer diagnostics and monitoring: Biomarkers



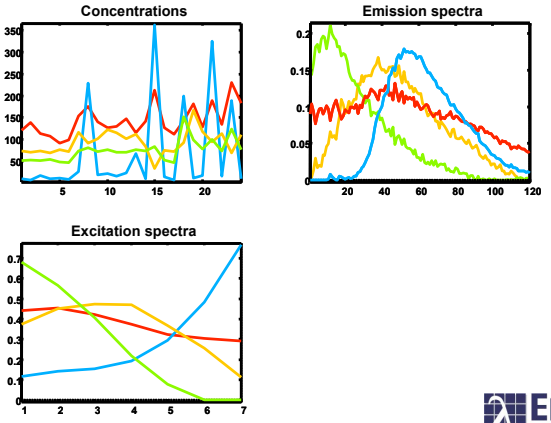
cancer

Cancer diagnostics

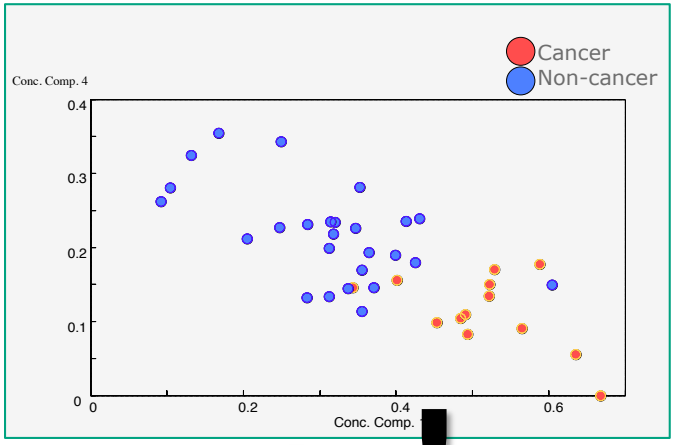
The three-way version
Here we just show the information obtained from a tiny part of the overall data



Cancer diagnostics



EIGENVECTOR RESEARCH INCORPORATED



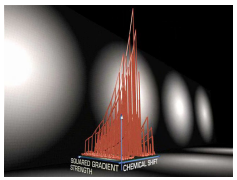
It works!

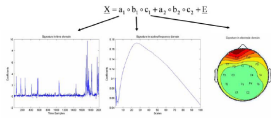
EIGENVECTOR RESEARCH INCORPORATED

Examples

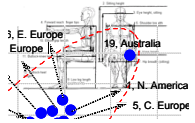
EEG

Metabonomics



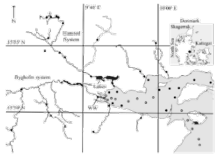


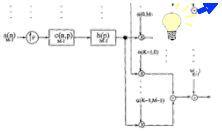
Anthropometry



Light-induced oxidation of cheese

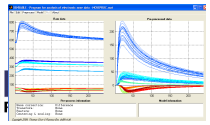
Tracing dissolved organic matter





Electronic nose

User separation in CDMA



EIGENVECTOR RESEARCH INCORPORATED

Using constraints

- **Example**
  - Instead of 'PCA': $\| \mathbf{X} - \mathbf{AB}' \|$
  - fit the model: $\| \mathbf{X} - \mathbf{AB}' \|$,

    subject to ***A*** *and* ***B*** *are nonnegative*

- Constraints are essential in two-way curve resolution because the model is unidentified

- In three-way curve resolution the model is often unique but constraints are still useful

EIGENVECTOR RESEARCH INCORPORATED

- **Obtain sensible parameters**
  - Ex.: Require chromatographic profiles to have but one peak
- **Obtain unique solution**
  - Ex.: Use selective channels in data to obtain uniqueness
- **Test hypotheses**
  - Ex.: Investigate if tryptophane is present in sample
- **Avoiding degeneracy and numerical problems**
  - Ex.: Enabling a PARAFAC model of data otherwise inappropriate for the model
- **Speed up algorithms**
  - Ex.: Use truncated bases to reexpress problem by a smaller problem
- **Enable quantitative analysis of qualitative data**
  - Ex.: Incorporate gender and job type predicting income

Why constraints?

EIGENVECTOR RESEARCH INCORPORATED

Aminoacid data

Try to model a three-component PARAFAC model of sample four and five – the two mixtures.

Does the model look good? If not; what to do?

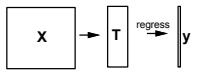
Try

EIGENVECTOR RESEARCH INCORPORATED

PLS for multi-way arrays

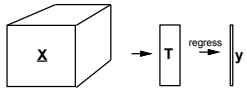
PLS

For two-way data a bilinear model is used



N-PLS

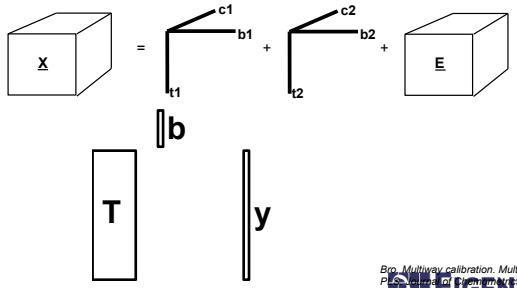
For three-way data a trilinear model is used



EIGENVECTOR RESEARCH INCORPORATED

Multilinear PLS regression

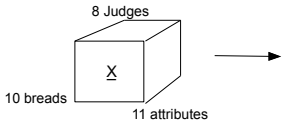
Use a trilinear (PARAFAC-like) model of **X** but such that the scores are predictive of **y**.



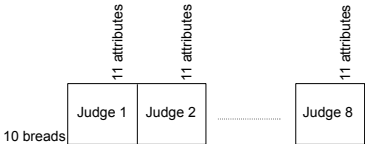
Bro. Multiway calibration. Multi-linear PLS. Journal of Chemometrics 10(1996) 47-61

EIGENVECTOR RESEARCH INCORPORATED

Three-way, two-way:
Does it make a difference?



5 breads (in replicates) × 11 attributes × 8 judges

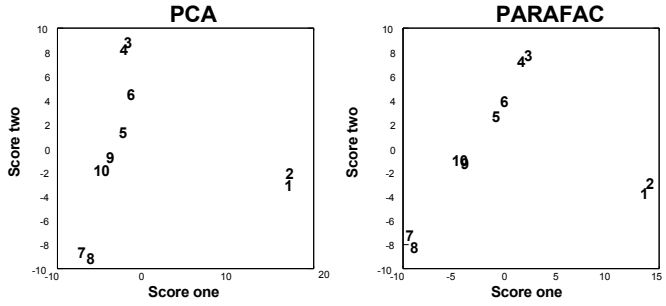


Sensory example

Data due to Magni Martens

EIGENVECTOR RESEARCH INCORPORATED

Scores PCA and PARAFAC

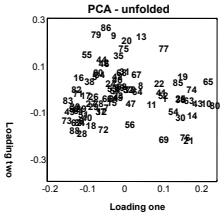


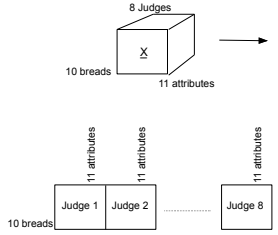
Similar but note that replicates are closer for PARAFAC

Three-way more robust because of 'stronger' structural model

EIGENVECTOR RESEARCH INCORPORATED

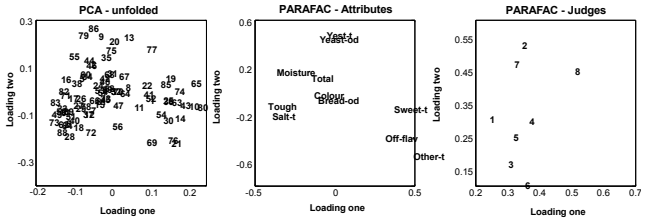
Loadings from bilinear PCA





EIGENVECTOR RESEARCH INCORPORATED

Loadings from PCA and PARAFAC



PARAFAC 19 loading-elements per component
PCA 88 loading-elements per component!

EIGENVECTOR RESEARCH INCORPORATED

SLPrandnu

SLPrantT

Calibration - predict salt content

EIGENVECTOR RESEARCH INCORPORATED

25% better with NPLS
Less overfit

	LV	Variation explained /%				RMSE	
		X cal.	X val.	Y cal.	Y val.	Y cal.	Y val.
SLPrandnu	1	43	25	80	62	0.21	0.29
	2	61	38	95	76	0.10	0.23
	3	74	49	100	84	0.03	0.19
SLPrantT	1	31	22	75	60	0.23	0.30
	2	46	36	93	82	0.12	0.20
	3	54	44	98	91	0.07	0.15

EIGENVECTOR RESEARCH INCORPORATED

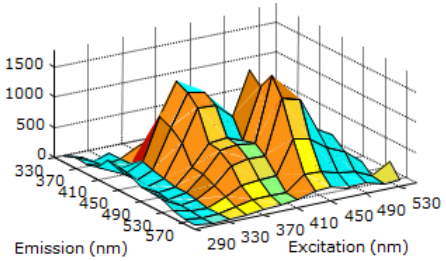
Example sugar processing

Sugar made from beets

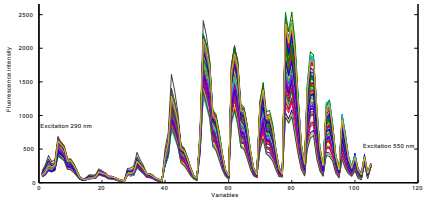
90 samples of white sugar measured by fluoroescence

EIGENVECTOR RESEARCH INCORPORATED

Example sugar processing



Three-way

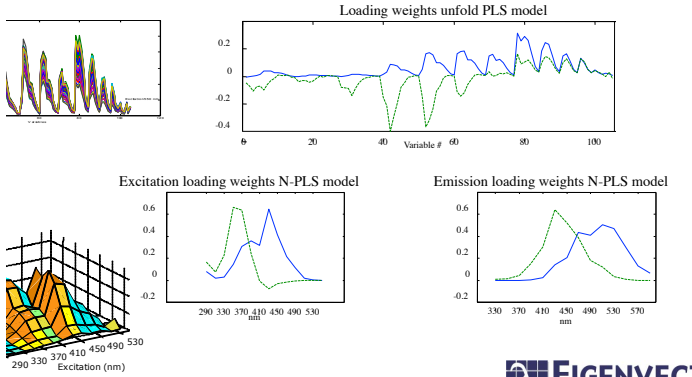


Two-way

EIGENVECTOR RESEARCH INCORPORATED

Trilinear model much more simple

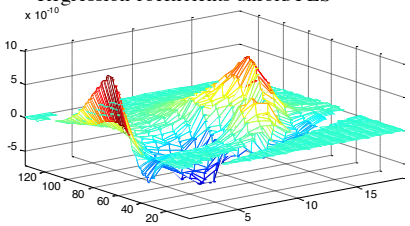
Hence easier to explore



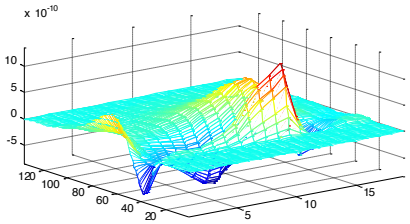
EIGENVECTOR RESEARCH INCORPORATED

The variance-stabilization of multi-way models

Regression coefficients unfold PLS



Regression coefficients multi-way PLS



EIGENVECTOR RESEARCH INCORPORATED

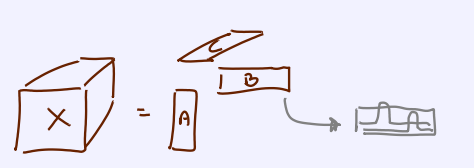
Important note on N-PLS

There is *NO* second order advantage in N-PLS

- You cannot handle new interferents that were not in the calibration set
- N-PLS works under the same premises as ordinary PLS

EIGENVECTOR RESEARCH INCORPORATED

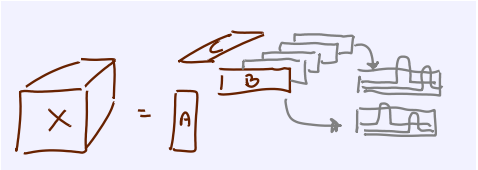
PARAFAC can not handle shifts and shape changes



PARAFAC(1) $\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^T$

PARAFAC2

PARAFAC2 for handling shifts*



**Actually it is more general than shifts but it's a feasible approximation to what PARAFAC2 can handle*

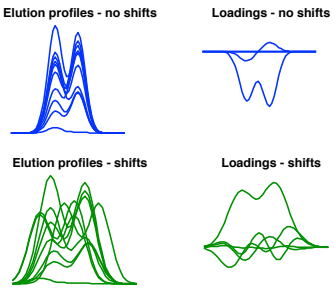
PARAFAC2 $\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}_k^T$ subject to $\mathbf{B}_k^T\mathbf{B}_k$ constant

PARAFAC(1) $\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}^T$

R. A. Harshman. *UCLA working papers in phonetics* 22:30-47, 1972.
H. A. L. Kiers, J. M. F. ten Berge, R. Bro. *J. Chemom.* 13:275-294, 1999.
R. Bro, C. A. Andersson, H. A. L. Kiers. *J. Chemom.* 13:295-309, 1999.

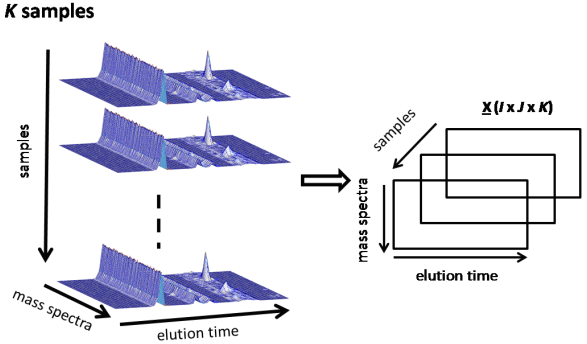
PARAFAC2 for shifted data

- **Two-way shifts**
  - Chromatography
  - Retention times constant => bilinear data
  - Retention times vary => breakdown

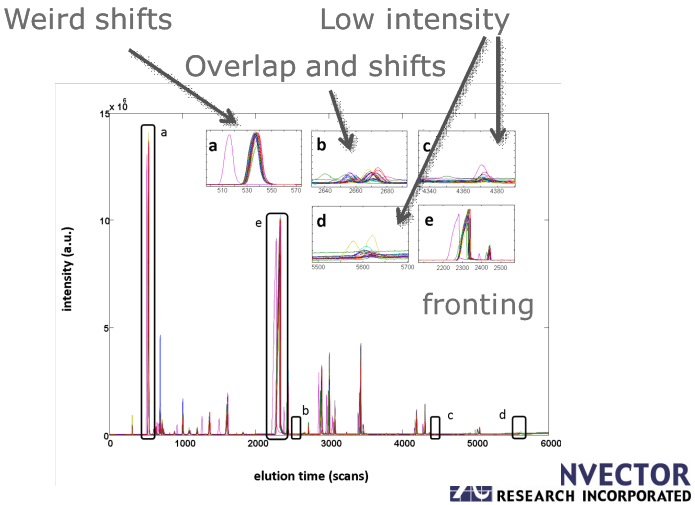


EIGENVECTOR RESEARCH INCORPORATED

60 wine samples measured by GC-MS

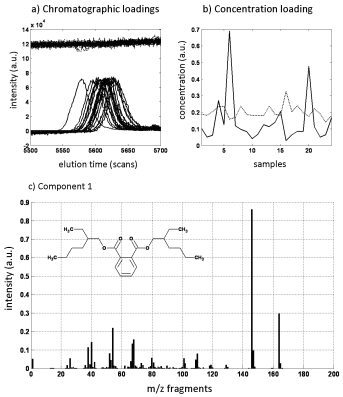


EIGENVECTOR RESEARCH INCORPORATED



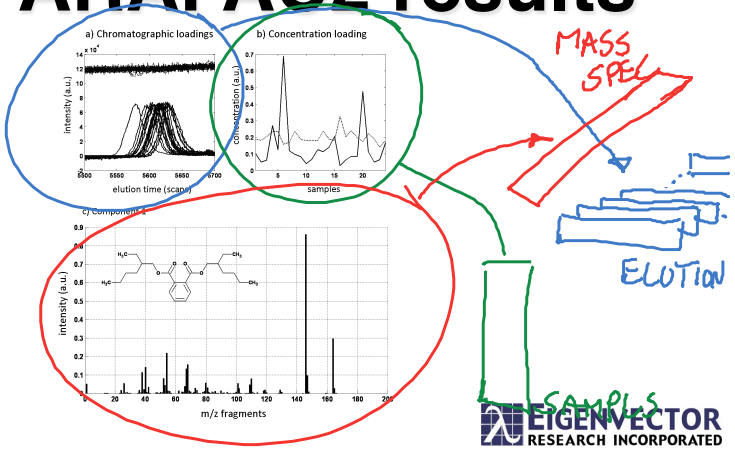
NVECTOR
RESEARCH INCORPORATED

PARAFAC2 results



EIGENVECTOR
RESEARCH INCORPORATED

PARAFAC2 results

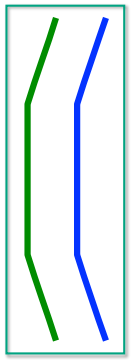


EIGENVECTOR
RESEARCH INCORPORATED

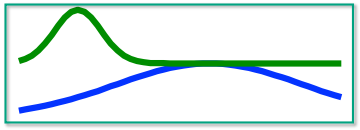
Tucker modeling

EIGENVECTOR
RESEARCH INCORPORATED

A (20[?]2) rank 1

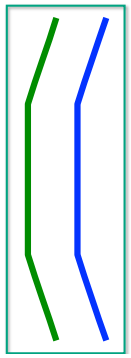


$\mathbf{B}^T$ (50[?]2) rank 2



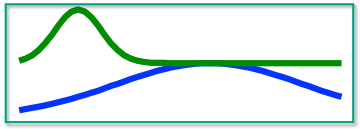
Row- and column ranks

A



$\mathbf{X}=\mathbf{A}\mathbf{B}^T$

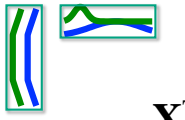
Rank 1 or 2?



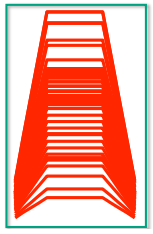
$\mathbf{B}^T$

Row- and column ranks

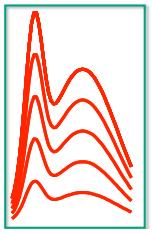
$\mathbf{X}=\mathbf{A}\mathbf{B}^T$ Rank 1 or 2?



$\mathbf{X}^T$



$\mathbf{X}$

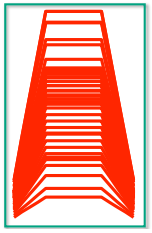


Row- and column ranks

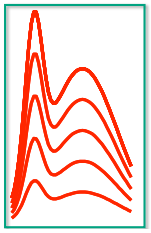
In ordinary algebra row = column = rank = 1.

This is boring!

$\mathbf{X}^T$



$\mathbf{X}$

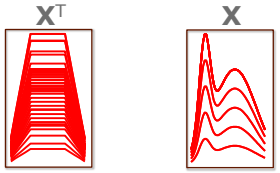


Row- and column ranks

In multi-way algebra

Row-rank ≠ column-rank ≠ rank.

Much more intuitive in fact – but highly unusual



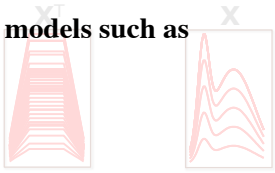
Row- and column ranks

In multi-way algebra

Row-rank ≠ column-rank ≠ rank.

Leads to subspace models such as Tucker3

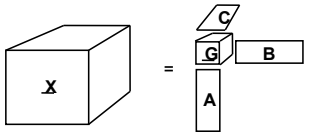
PARAFAC



Row- and column ranks

The Tucker3 model

- For three-way data, three orthogonal bases, **A**, **B**, and **C**; one for each mode
- Tucker3 is **X** = **AG**(**C** ⊗ **B**)' + **E**
- Loadings are truncated bases and **G** the representation of **X** in these reduced spaces



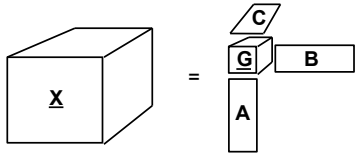
L. R. Tucker. The extension of factor analysis to three-dimensional matrices. In: *Contributions to Math. Psychology*, Eds. Frederiksen, Gulliksen, New York:Holt, Rinehart & Winston, 1964

L. R. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika* 31:279-311, 1966

EIGENVECTOR RESEARCH INCORPORATED

Differences from PARAFAC:

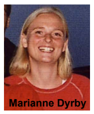
- The number of components can vary in **A**, **B**, and **C**!
- **G** is not superdiagonal
- Tucker loadings not unique (only subspace) = rotational freedom
- Tucker loadings orthogonal => variance-partitioning

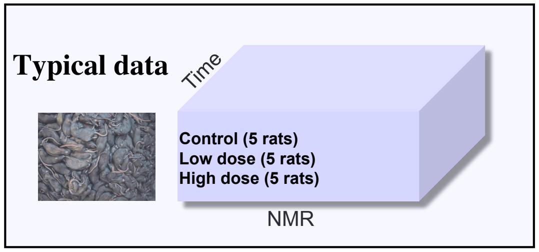


Tucker3 versus PARAFAC

EIGENVECTOR RESEARCH INCORPORATED

Toxic study





ELSEVIER Chemometrics and Intelligent Laboratory Systems 76 (2005) 79–89
www.elsevier.com/locate/chemolab

Multiway chemometric analysis of the metabolic response to toxins monitored by NMR

Marianne Dyrby[a], Dorrit Baunsgaard[b], Rasmus

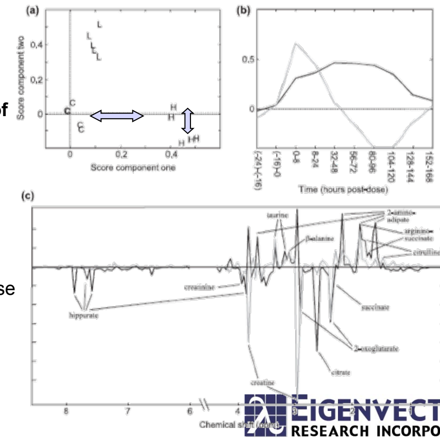
EIGENVECTOR RESEARCH INCORPORATED

Toxic study

Hydrazine study

Benefit of Tucker3 instead of commonly used unfolding:
Shows recovery of low-dose rats as well as an early response to taurine and creatine.

Remaining problem:
Are components reflecting dose effect or biological variation within dose?



EIGENVECTOR RESEARCH INCORPORATED

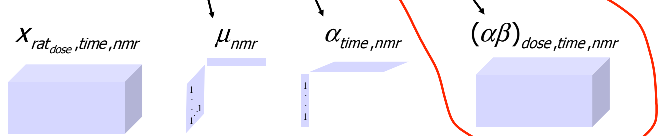
Toxic study

•Further analysis
•Use ANOVA-SCA (simultaneous component analysis)

ANOVA-Simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data
AK Smilde, JJ Jansen, HCJ Hoefsloot, RN Lamers, J van der Greef, ME Timmerman

ANOVA

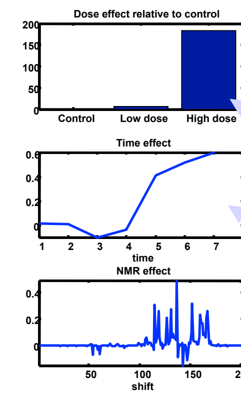
$$x_{rat_{dose},time,nmr} = \mu_{nmr} + \alpha_{time,nmr} + (\alpha\beta)_{dose,time,nmr} + e_{rat_{dose},time,nmr}$$



EIGENVECTOR RESEARCH INCORPORATED

•PARAFAC on dose/time-effect

Separate the effect into a shock

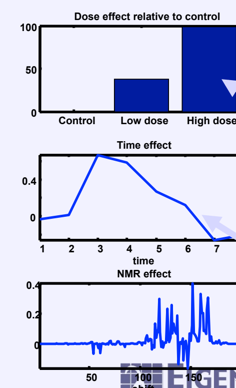


High-dose only

Irreversible

PARAFAC factor 1

and a reversible effect



All-dose linear

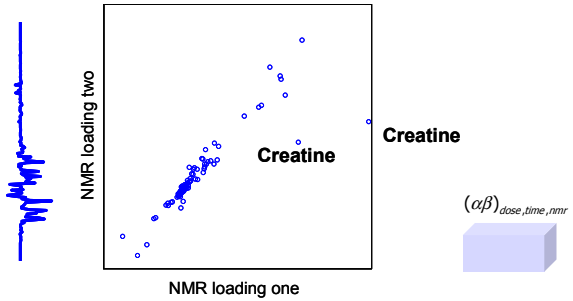
Reversible

PARAFAC factor 2

EIGENVECTOR RESEARCH INCORPORATED

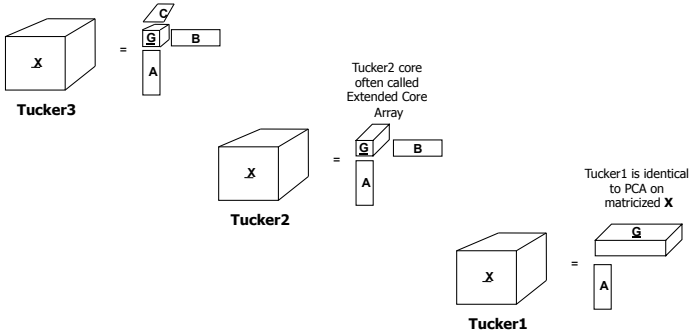
Toxic study

Difference between reversible and irreversible effect
Creatine indicating chronic kidney damage



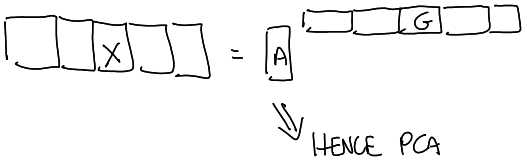
EIGENVECTOR RESEARCH INCORPORATED

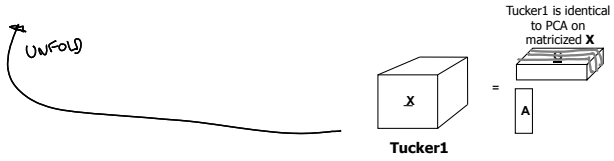
Tucker3 has the number 3 because three modes are 'reduced'.
Tucker2 and Tucker1 reduces two and one modes respectively



Other Tucker models

EIGENVECTOR RESEARCH INCORPORATED



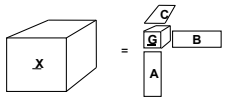


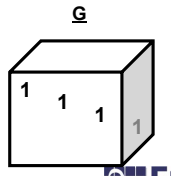
Other Tucker models

EIGENVECTOR RESEARCH INCORPORATED

NB: Core consistency

PARAFAC Can be written as a constrained Tucker3 model

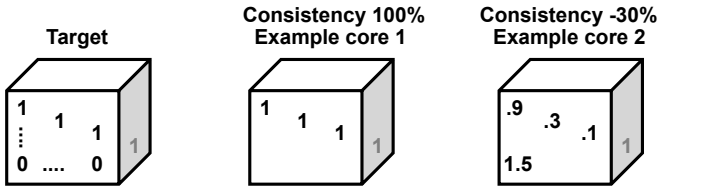




EIGENVECTOR RESEARCH INCORPORATED

Core consistency

Test PARAFAC model using PARAFAC **A**, **B** and **C** and see what an unconstrained core gives
I.e. the core = **X** "divided" by **A**, **B** and **C**

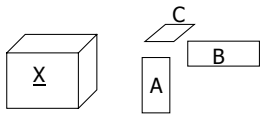


EIGENVECTOR RESEARCH INCORPORATED

Cook book

EIGENVECTOR RESEARCH INCORPORATED

- **Principle**
  - Fit PARAFAC model with 1 to F components
  - Calculate core consistency
    - –=% of trilinear variation in the model space
  - If << 100%, wrong # components



How to

EIGENVECTOR RESEARCH INCORPORATED

- **How to fit multi-way models**
  - Everything said below is of inferior importance compared to knowing your data and the models you use and that you use the models *very* critically
- **How to fit N-PLS**
  - Exactly as two-way PLS (cross-validation etc.)
- **How to fit PARAFAC/Tucker**
  - Can be more cumbersome ….

In practice?

EIGENVECTOR RESEARCH INCORPORATED

- **How to fit PARAFAC/Tucker**
  - Screen raw data etc. and deal with extreme outliers
  - Do initial PCA models on different two-way versions
    - Note potential outliers for later
    - Note the rank in each mode (points to possible rank of three-way model). If rank is *P*, *Q*, *R* of the three matricized arrays, then a (*P*, *Q*, *R*) Tucker3 *will* do the job. PARAFAC may also be applicable even though the ranks are different.
  - Do initial PARAFAC/Tucker3 models
    - Use appr. correct number of components as experienced from above (several alternative ones)
    - Explore explained variance compared to noise level, explore loadings, scores, residuals to find indications of too many or too few components being used
    - For PARAFAC note indications of too many components (many iterations needed, low core consistency, local minima etc.)

In practice?

VECTOR RESEARCH INCORPORATED

… Choice between PARAFAC and Tucker

- Multi-way and multi-linear is not the same thing
- Any multi-way dataset can be modeled with Tucker.
  - If the ranks are low, this is feasible
  - Argument similar to PCA on two-way data
  - *Tucker is almost as parsimonious as PARAFAC compared to PCA*
- Only some datasets can be modeled with PARAFAC
  - If the data approximately follow the multilinear model of PARAFAC
  - Hence, PARAFAC when *a priori* tells so or when uniqueness is desperately needed and Tucker otherwise

In practice?

TOR RESEARCH INCORPORATED

Some attention required when interpreting Tucker3 models

- In PARAFAC, PCA etc. each score/loading is only involved in one component
- In Tucker3 all interactions are allowed
- Therefore, bi-plots can not be made immediately

- Interpret loadings from one mode at a time
- Combine only when taking the core values into account

Interpreting Tucker3

Number of components?

- **PARAFAC**
  - Simpler than PCA (but takes more time):
  - Cross-val, Scree etc. as in PCA

Mainly these →

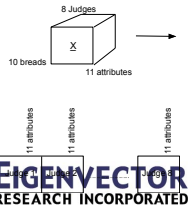
  - Core consistency
  - Plus split-half
  - Plus algorithmic indications (degeneracy, many iterations, local minima etc.)
- **N-PLS**
  - As in bilinear PLS
- **Tucker**
  - Tough one, but basically as in PCA, except there are now three numbers of components to choose

EIGENVECTOR RESEARCH INCORPORATED

Cross-validation

- Cross-validation hardly ever used for PARAFAC and Tucker
- Below: PARAFAC fits worse but provide best predictions
- Thus nothing gained going to more complex Tucker3 or even more complex PCA (Tucker1)
- Note that PCA fits indicates that PCA is excellent!

Number of components	PARAFAC		Tucker3		Tucker1	
	Fit	*Cross-val*	*Fit*	*Cross-val*	*Fit*	*Cross-val*
1	35.3	14.5	35.3	14.5	44.6	13.2
2	49.2	26.2	49.2	26.2	65.8	26.5
3	57.4	32.9	57.7	31.6	74.3	18.6
4	62.7	34.4	64.6	19.6	80.7	<0
5	67.2	33.0	72.7	24.6	86.2	<0



EIGENVECTOR RESEARCH INCORPORATED

Using the models

- **PARAFAC (& PARAFAC2)**
  - Algorithm occasionally slow & problematic
  - Requires some experience
  - Not nested
  - Unique
  - Solutions easy to interpret
- **N-PLS**
  - Algorithm fast & robust
  - Non-'unique'
  - Nested
  - Solutions easy to interpret
- **Tucker**
  - Algorithm fast & robust
  - Non-unique
  - Not nested
  - Solutions difficult to interpret

EIGENVECTOR RESEARCH INCORPORATED

Conclusion

Tensor models provide

Mathematical chromatography
Huge noise reduction
Intuitive models (chemically)
Better handling of correlations
Simpler interpretation
Robustness

EIGENVECTOR RESEARCH INCORPORATED