



Applications of Tensor Methods in Life Sciences

Rasmus Bro

University of Copenhagen
Faculty of Life Sciences
rb@life.ku.dk



PARAFAC

A very nice model

Some examples

How to store a cheese?

A model of wine

Some issues

Variable selection

Nonnegativity

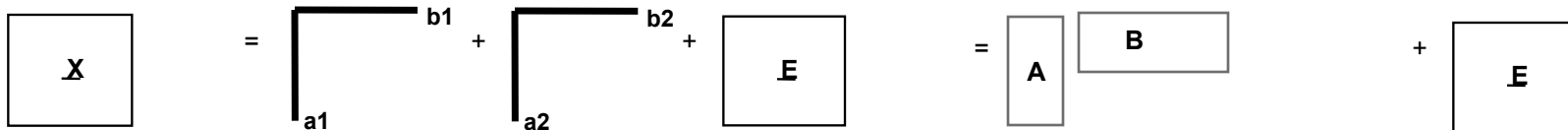
Dealing with missing data



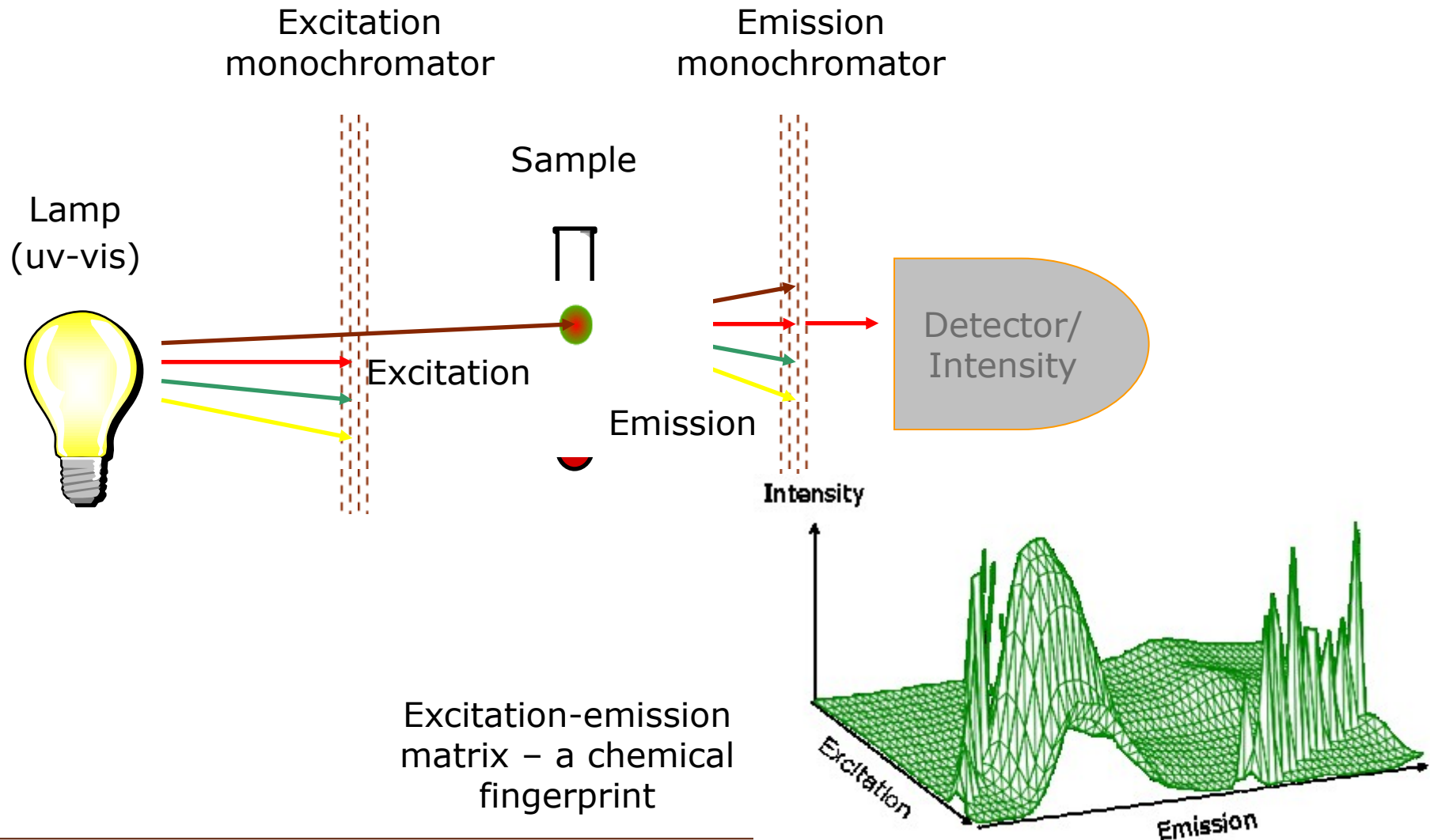
PARallel FACTor analysis

- PCA - bilinear model,

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

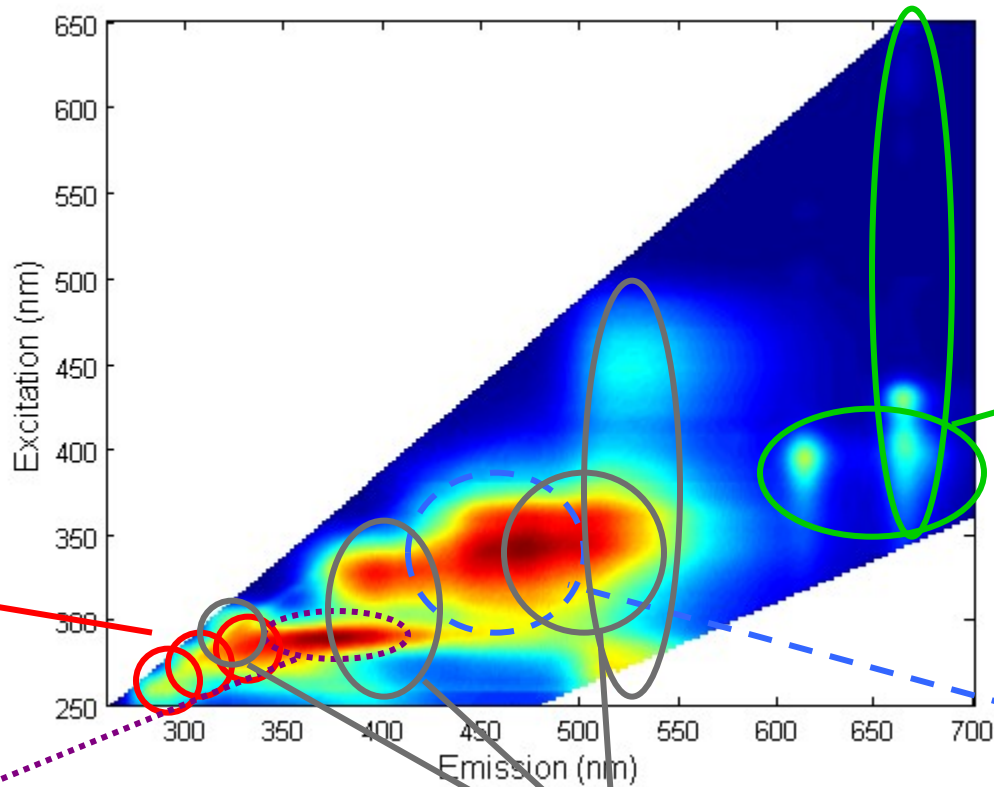


Fluorescence spectroscopy



Fluorescence excitation-emission

Very high sensitivity and selectivity towards important compounds



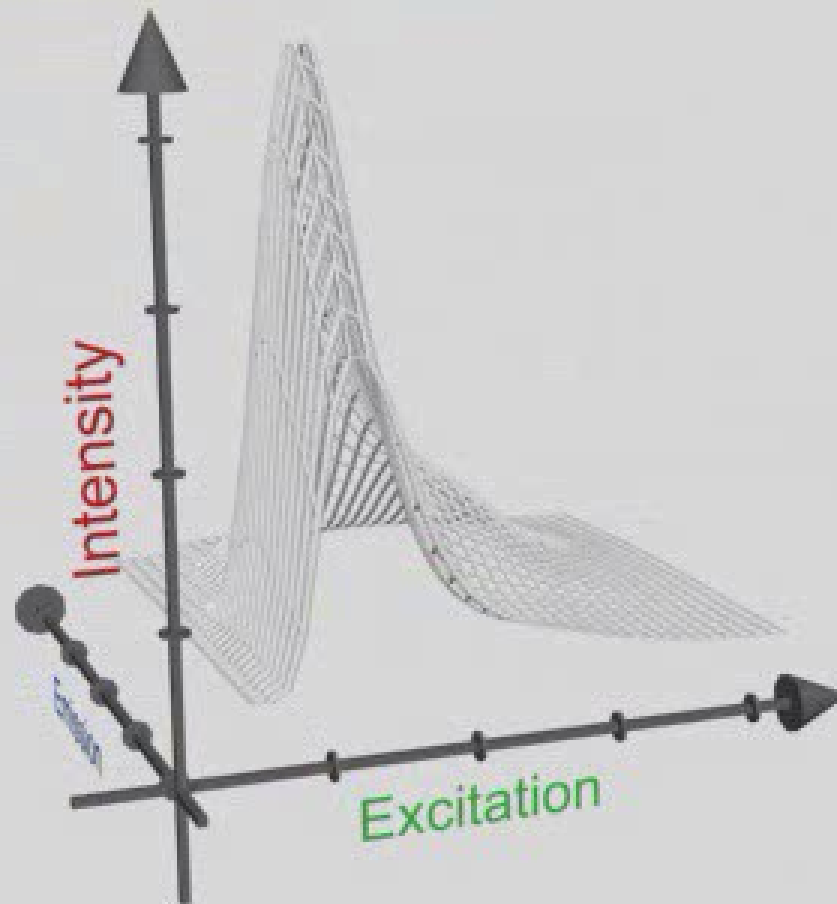
Chlorophyll
Porphyrin

Amino acids

NADH

ATP

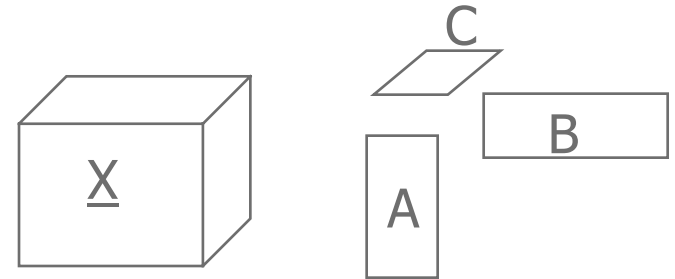
Vitamins
(A, B2, B6, E)



PARAFAC - uniqueness

- **No rotational freedom**

Unlike the bilinear 'PCA' model, there is only *one* solution



- **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 be 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.



PARAFAC is mathematical chromatography

aka

Blind source separation

Solving the cocktail party effect

Unmixing

Curve resolution

...

Mathematical chromatography eliminates major problems in multivariate analysis:

- Indirect correlations stemming from rotational freedom
- It also eliminates outliers
- It determines underlying sources
- Simpler because it provides a chemical model
- It is way more noise insensitive



How to store a chees

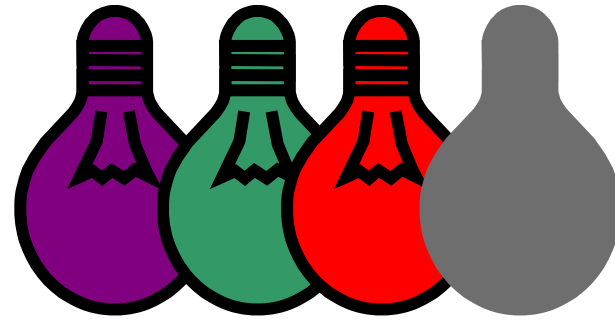
Oxidation

- Oxidation from light causes rancid taste of cheese, butter etc.
- Important for packaging of food and shelf-storage
- Believed to be caused by riboflavin acting as photosensitizer
- Riboflavin does not absorb much red light, hence red material should protect



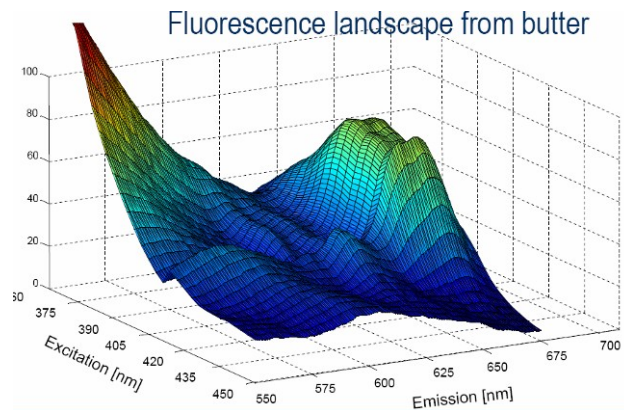
Experiment

Different light
With / without Oxygen
Different storage time



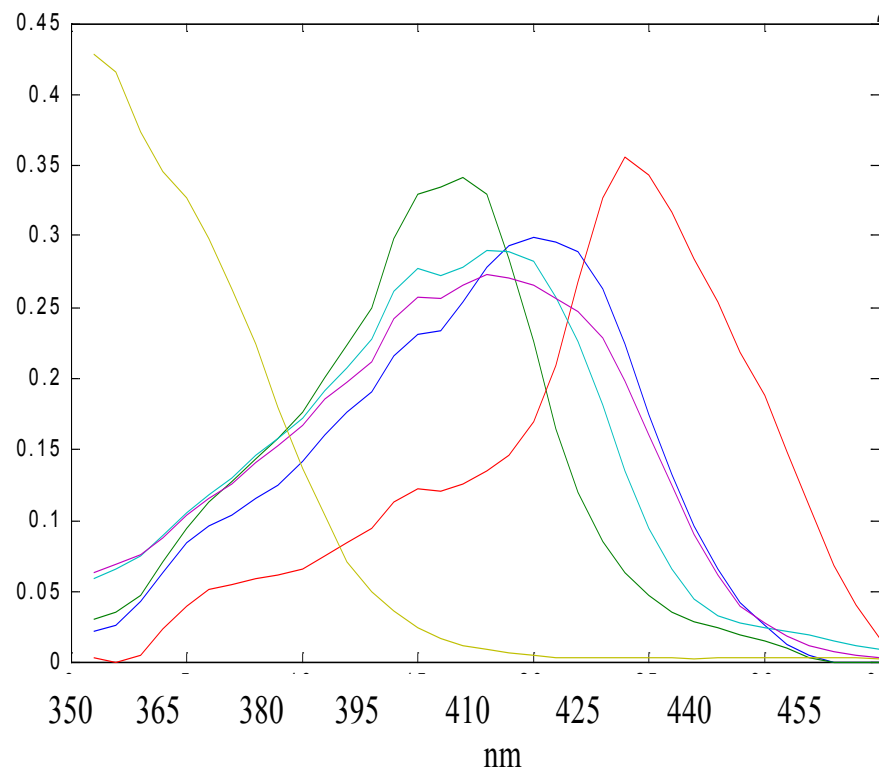
Samples measured by

- Sensory analysis (quality)
- Fluorescence EEM

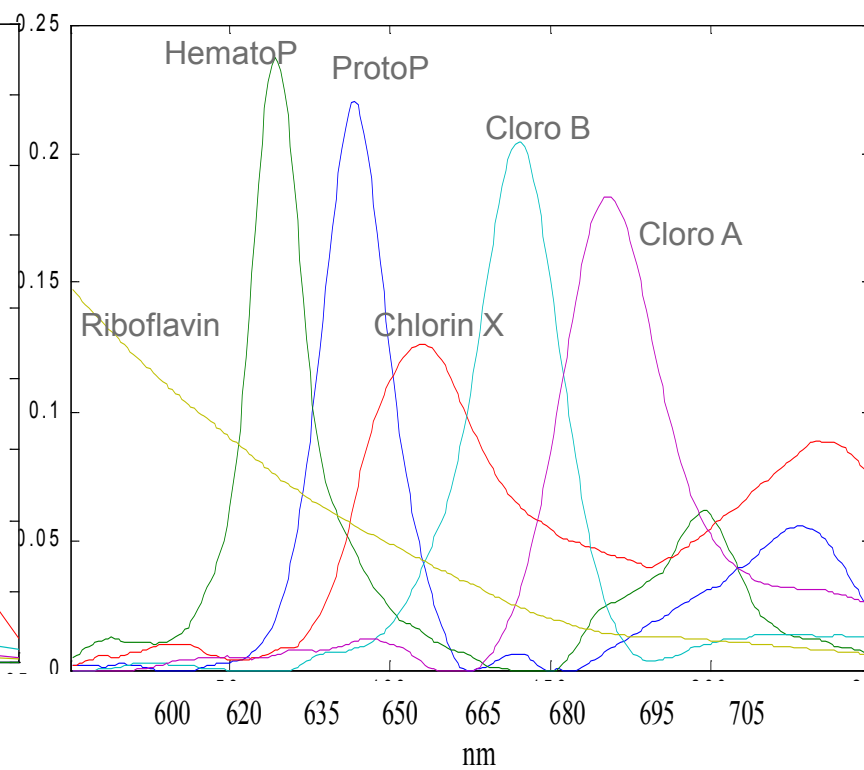


Spectra from PARAFAC of EEMs

Excitation



Emission

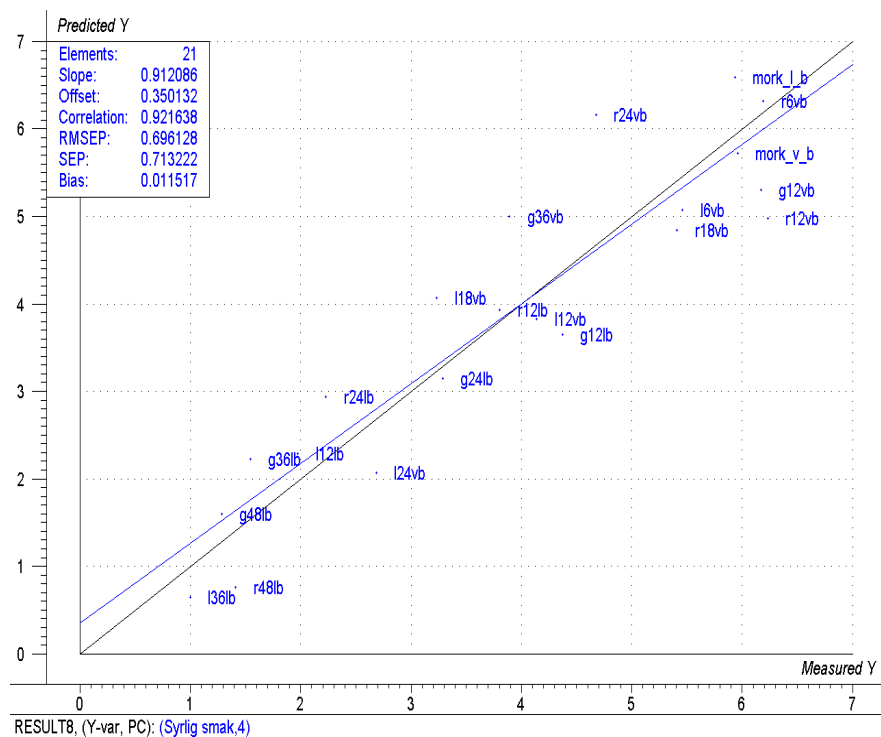


From JPWo/Matforsk

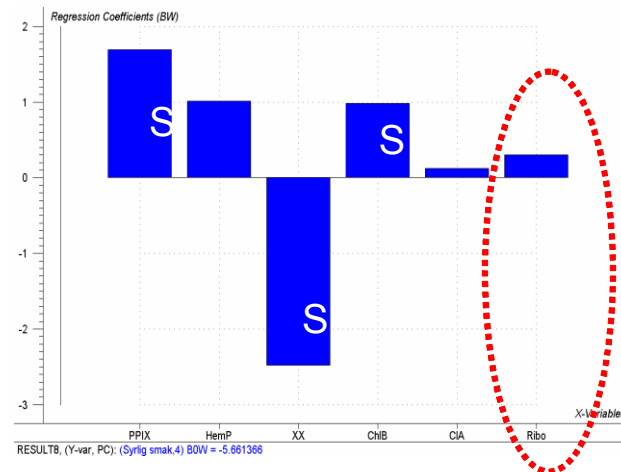


Relation between sensory data and PARAFAC estimated concentrations

Rancid taste



Importance of different compounds



Protoporphyrin
Chlorophyll B
X (Chlorine?)

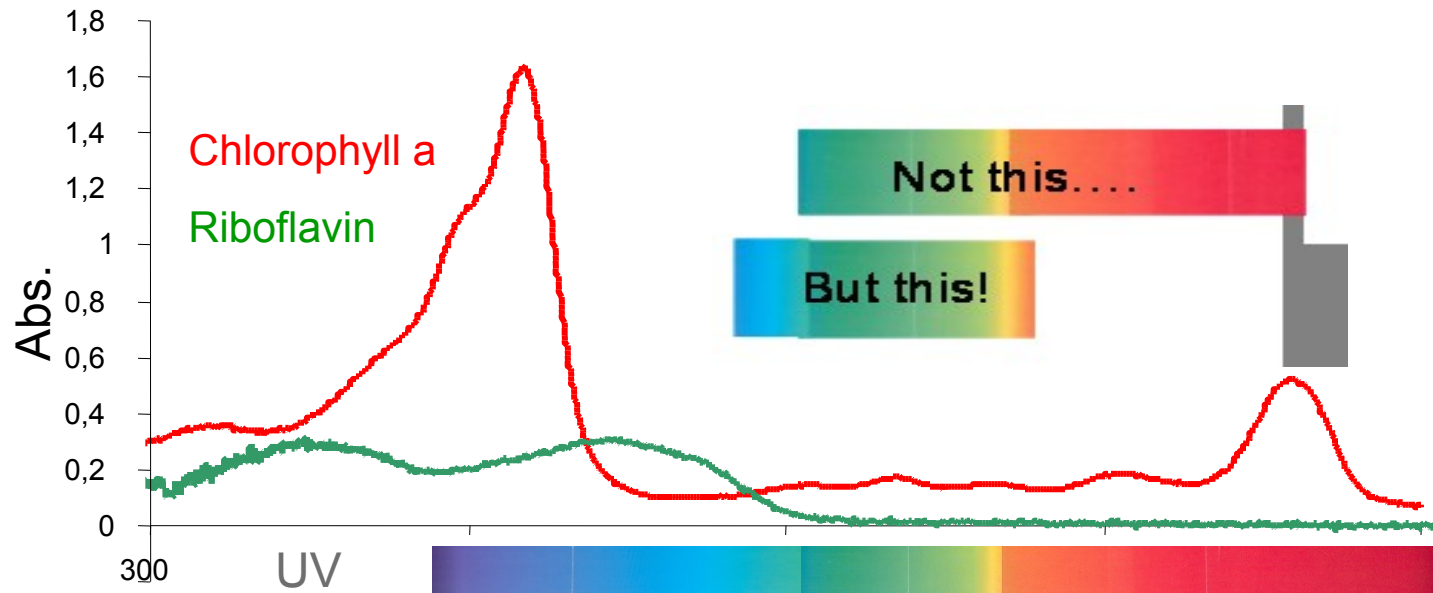
From JPWo/ Matforsk

From JPWo/Matforsk



New result

- Apart from riboflavin at least five other light-sensitizers
- 'New' ones seem to be more important than riboflavin
- Fluorescence and PARAFAC provides a 'simple' approach for exploring these.



From JPWo/Matforsk



A wine model

PARAFAC can not handle shifts and shape changes



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



A wine model

PARAFAC can not handle shifts and shape changes



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



PARAFAC2

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



PARAFAC2

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

PARAFAC(1)

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

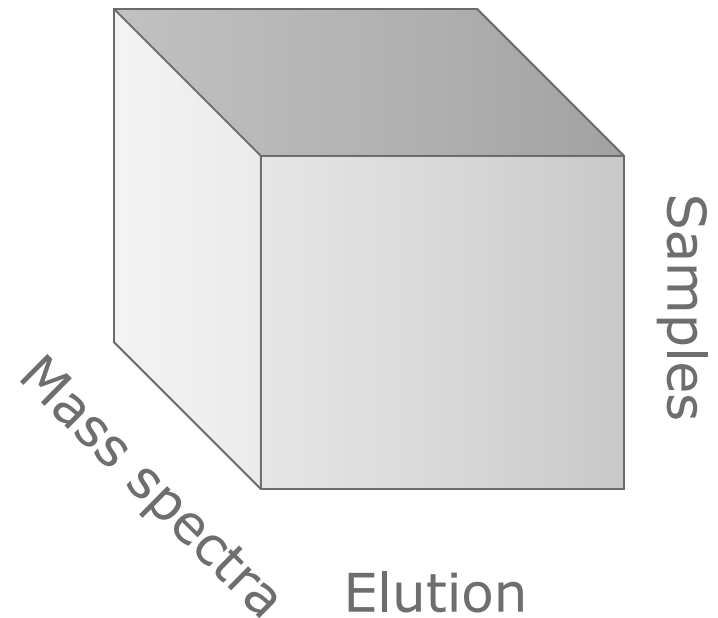
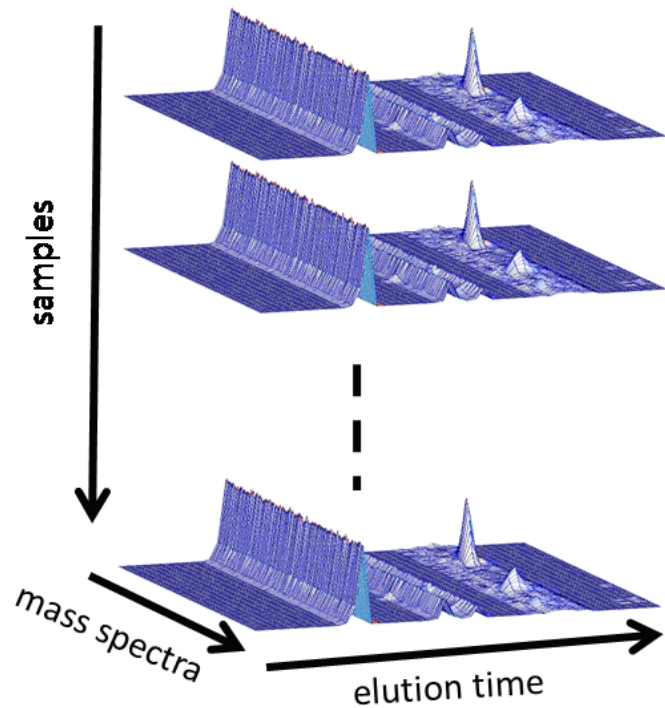
R. A. Harshman. *UCLA working papers in phonetics* 22:30-47, 1972.

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.

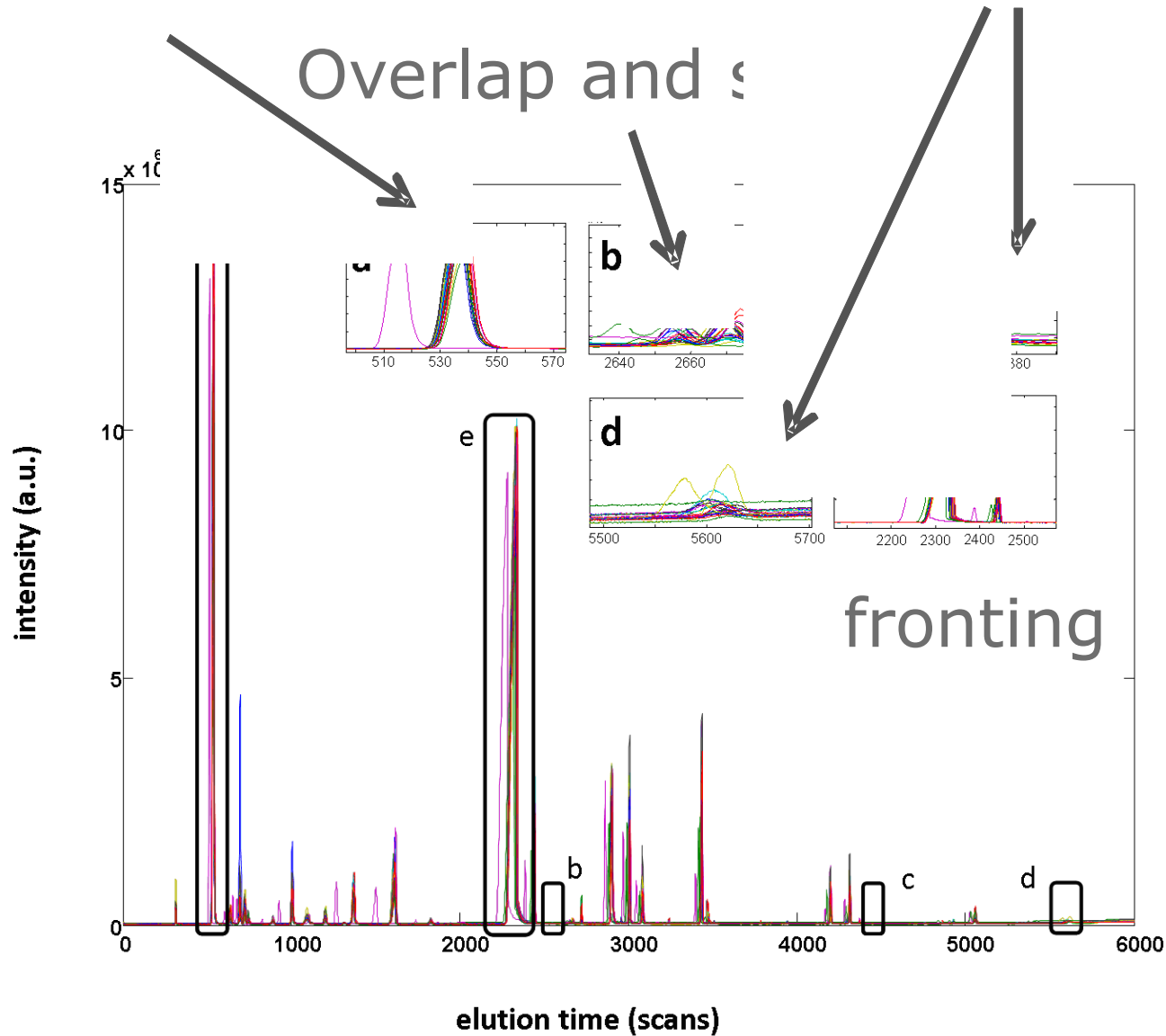


60 wine samples measured by GC-MS



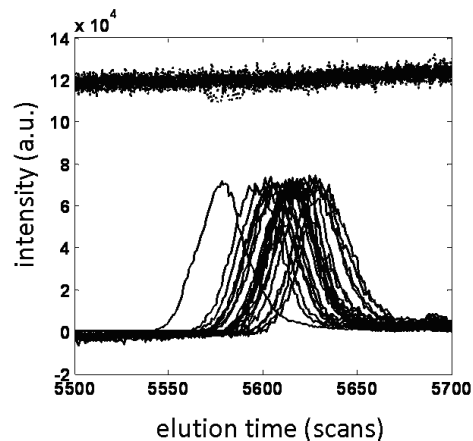
Weird shifts

Low intensity and baseline

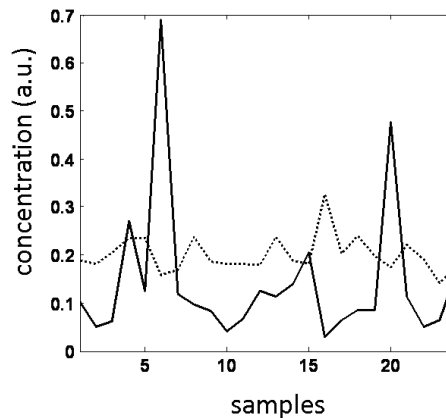


PARAFAC2 results

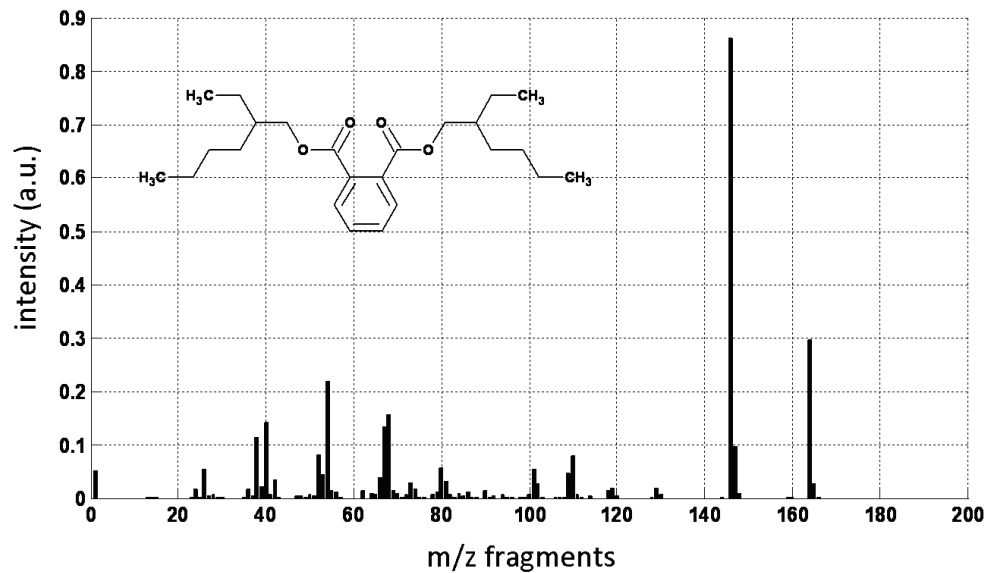
a) Chromatographic loadings



b) Concentration loading



c) Component 1



Other applications of tensor methods

Scientific field

Environmental monitoring

Sensory analysis

Process monitoring

Fermentation

Cell phone audio quality

Wireless communication

Metabolomics

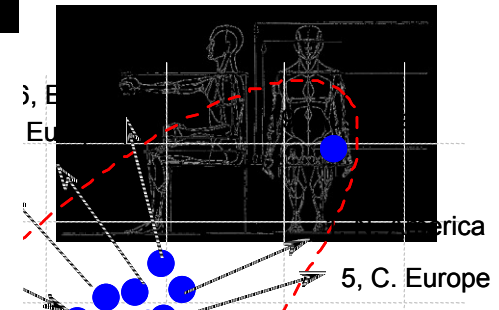
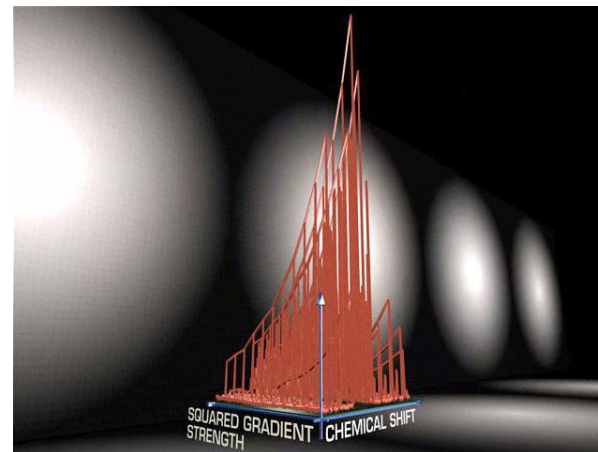
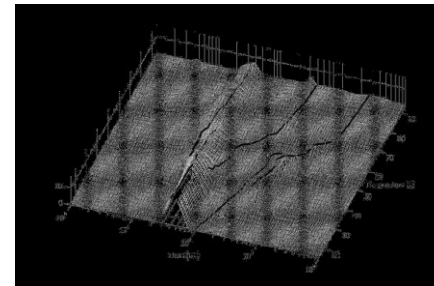
Proteomics

Cancer diagnostics

Anthropometry

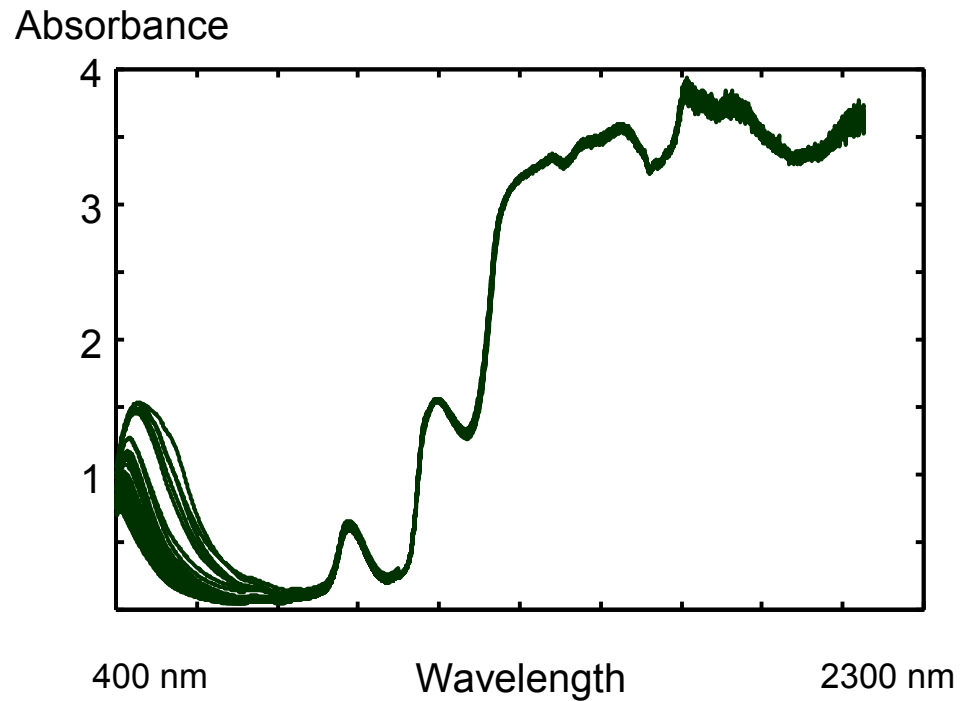
...

...

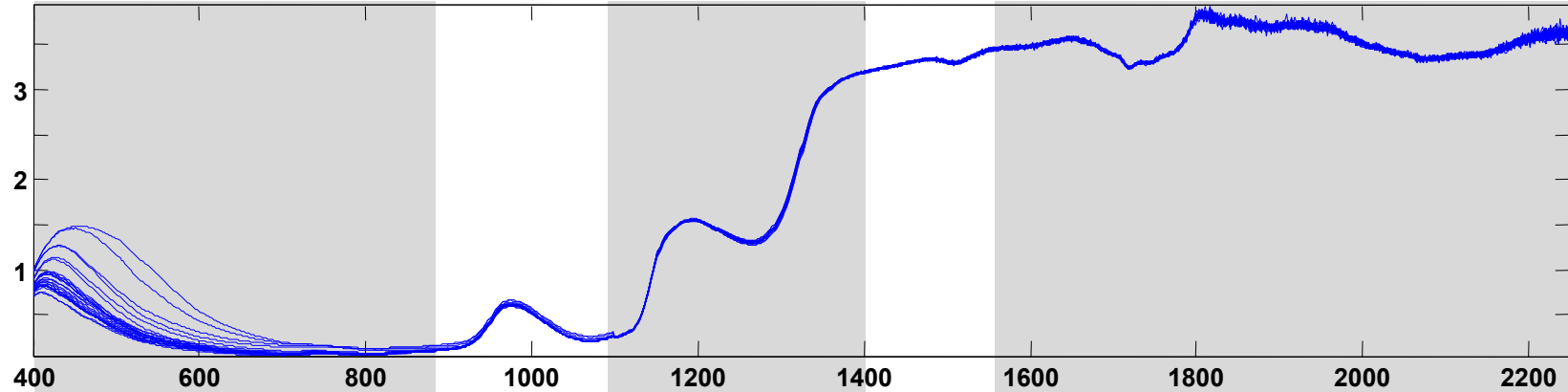


Variable selection

VIS/NIR spectra of 61 beers



Raw data

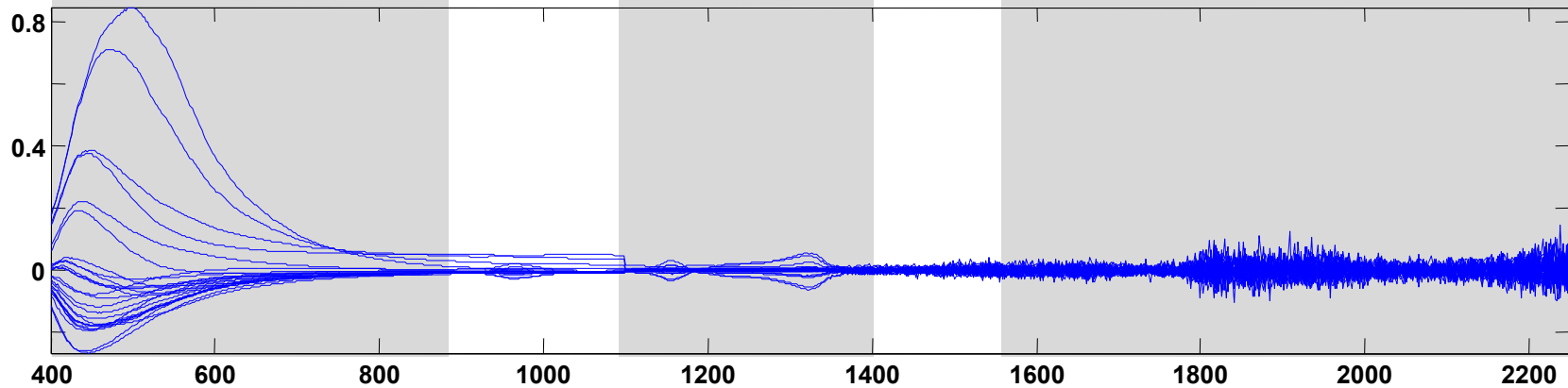


Not relevant
but highly
systematic

The
good
part

Just crap! Random
noise

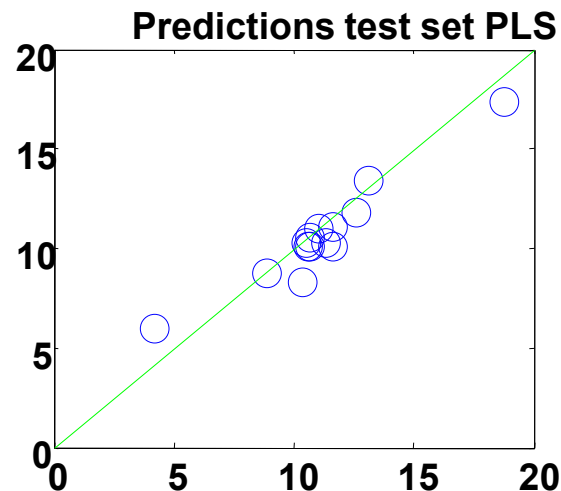
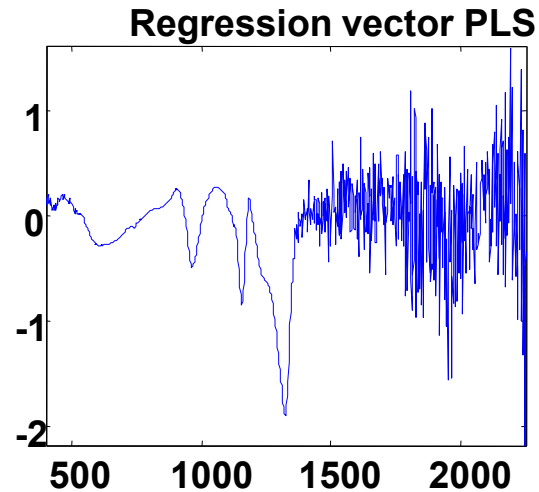
Actual modelled data - centered



'Classical' regression – in this case partial least squares (PLS). Good!

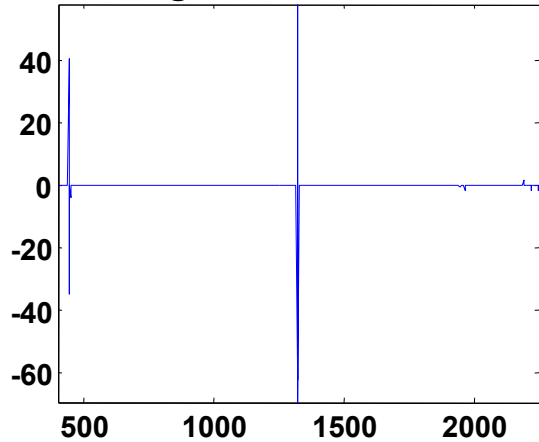
And can be optimized by chemical interpretation

But – this is not always the case

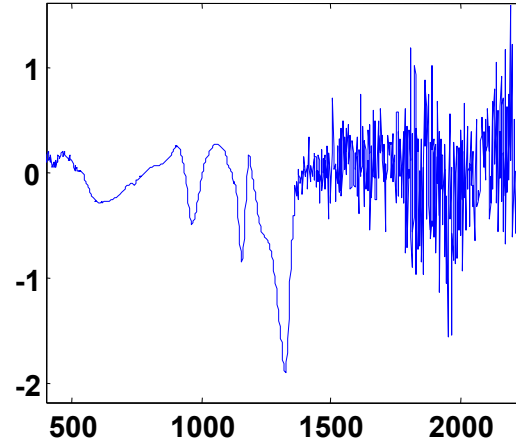


Lasso. Weird stuff! Important area represented by two variables.

Regression vector LASSO

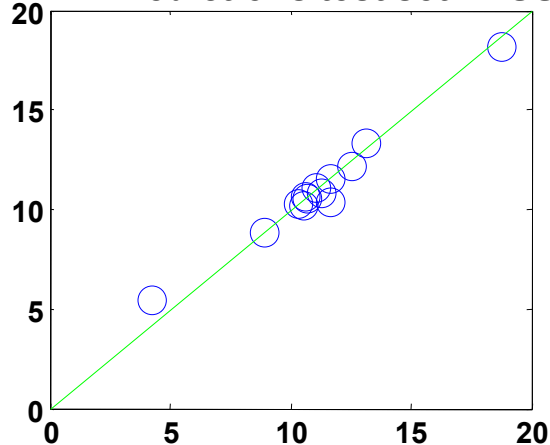


Regression vector PLS

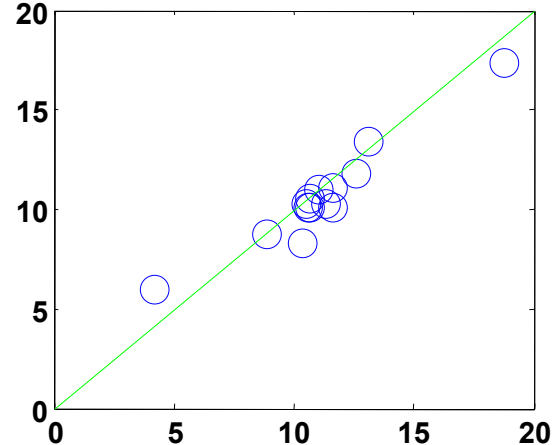


Little support:
Fragile
Non-robust
Poor outlier ability
Interpretation low

Predictions test set LASSO



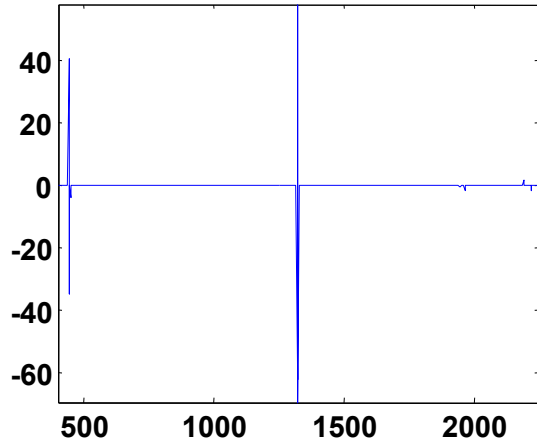
Predictions test set PLS



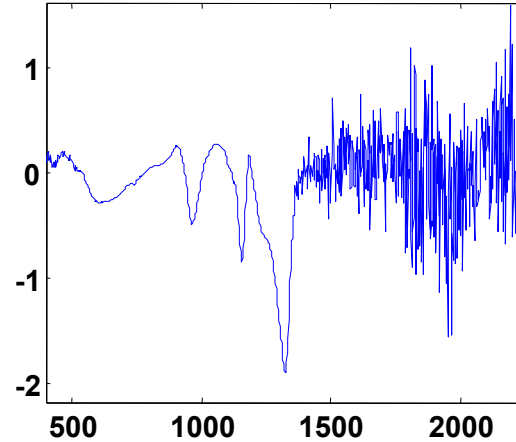
'Chemometric' variable selection – very nice!

Variable selection by selectivity ratios but others would do the job as well

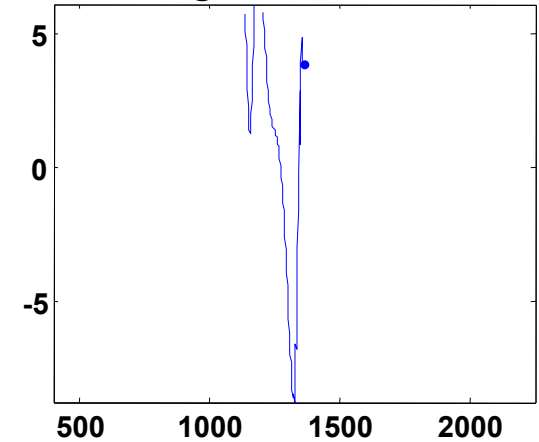
Regression vector LASSO



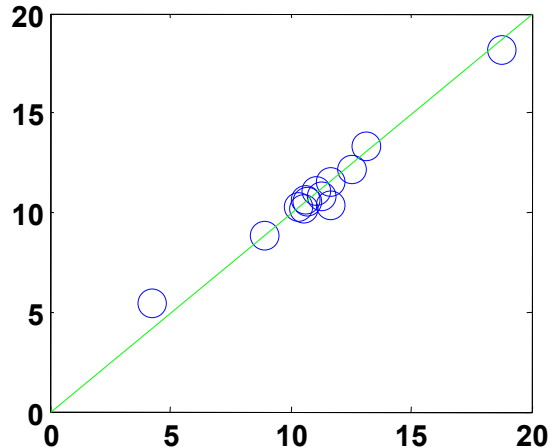
Regression vector PLS



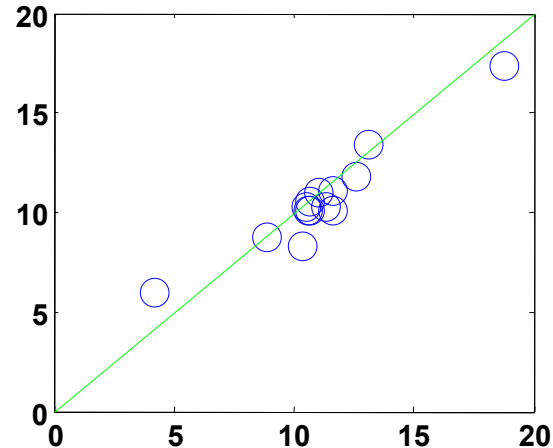
Regression vector SR



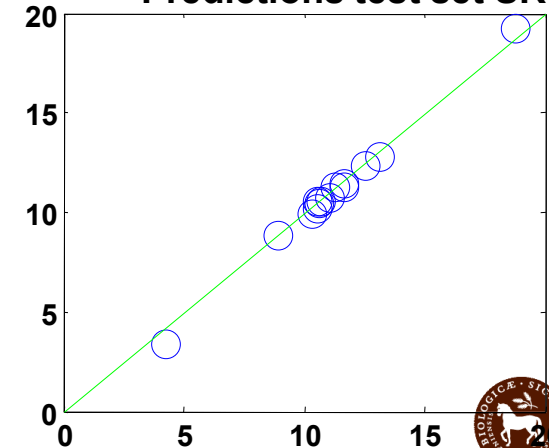
Predictions test set LASSO



Predictions test set PLS



Predictions test set SR



Nonnegativity

In 1999 Lee and Seung wrote a Nature paper on NMF - non-negative matrix factorization.

However, NMF has existed for much more than 30 years under the name multivariate curve resolution.

Classical papers related to NMF

Lawton & Sylvestre. Self modeling curve resolution. *Technometrics* 13:617-633, 1971.

Hanson & Lawson. *Solving least squares problems*, Englewood Cliffs:Prentice-Hall, Inc, 1974.



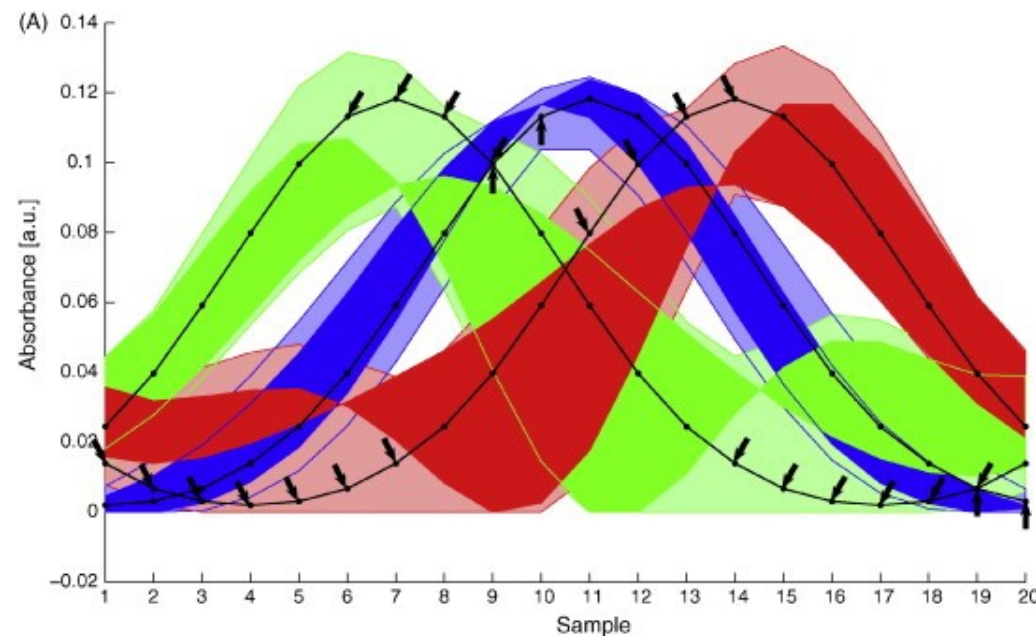
Some facts worth noting

NMF is not generally unique – rotational freedom

Conditions for uniqueness exist

When unique, the solution is often shaky

NMF is (very) sensitive to starting values



Dealing with missing data

No missing

Ex.: standard PCA loss function $||\mathbf{X}-\mathbf{TP}'|| = \sum_{i=1}^I \sum_{j=1}^J \left(x_{ij} - \sum_{f=1}^F t_{if} p_{jf} \right)^2$

I.e., a summation of errors over all elements of \mathbf{X}



How can that loss function be optimized?

Method 1: use weighted least squares regression

Method 2: use imputation (expectation maximization)

1. Put numbers in missing elements
2. Fit model to these 'wrong' data (Ex: $\mathbf{M} = \mathbf{TP}$ ' in PCA)
3. Replace missing elements with model guess (Ex: $x_{ij} = M_{ij}$ in PCA)
4. Go to step 2 until convergence

Both methods give same result. Method 2 is easy to implement, Method 1 sometimes faster, but more memory-demanding



Some concluding remarks

Tensor models provide

Mathematical chromatography (real blind source separation)

Huge noise reduction

Intuitive models (chemically)

Better handling of correlations

Robustness

...

But you need to know your data well – or be lucky

Much needed

Better algorithms

Better statistical diagnostics

Better software



Papers, m-files, courses, database of references, data sets, spectral libraries etc.

www.models.life.ku.dk

Rasmus Bro
rb@life.ku.dk