TENSOR DECOMPOSITIONS

SOLVING FUNDAMENTAL PROBLEMS IN CHEMISTRY

Rasmus Bro
rb@life.ku.dk

Outline

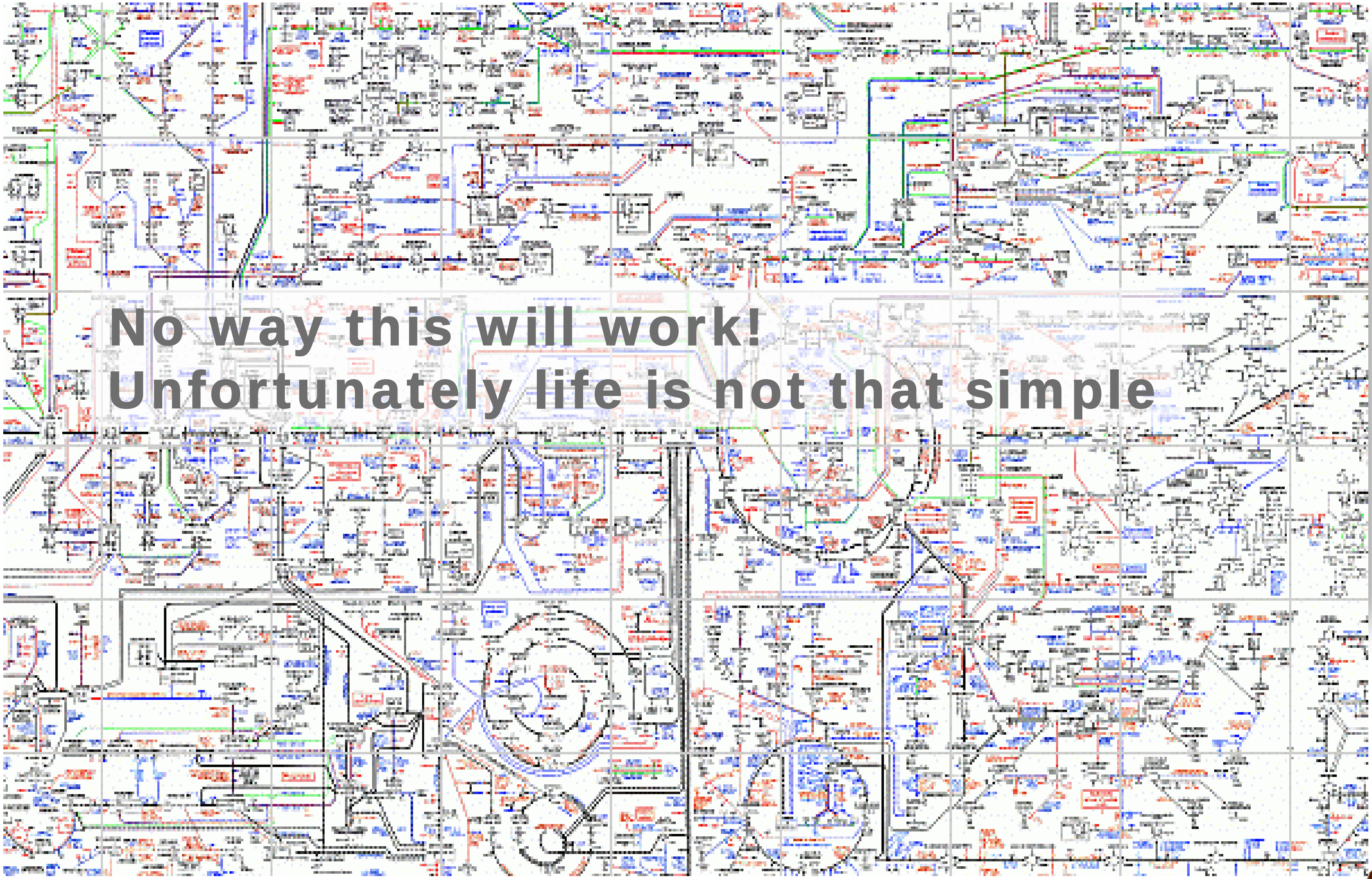
Limitations in univariate analysis

Limitations in multivariate analysis

Tensor models – mathematical chromatography

Typical quest in systems biology and omics:

Find *the* cancer marker (or something to that effect)



No way this will work!

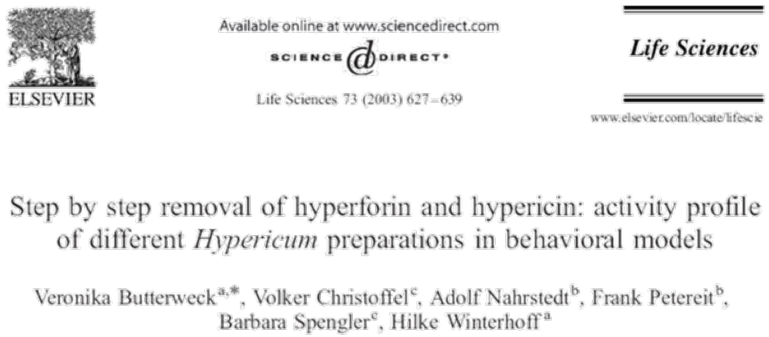
Unfortunately life is not that simple

Example from plant medicine: hypericum

The mechanism as an antidepressant not fully understood but originally thought to be due solely to hypericin

Official regulations require:

H. perforatum standardized to contain 0.3% hypericin



ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT

Life Sciences 73 (2003) 627–639

Life Sciences

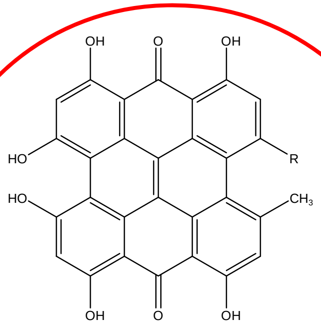
www.elsevier.com/locate/lifescie

Step by step removal of hyperforin and hypericin: activity profile of different *Hypericum* preparations in behavioral models

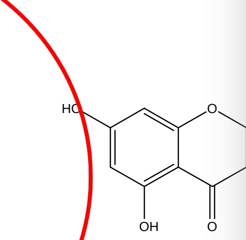
Veronika Butterweck[a,*], Volker Christoffel[c], Adolf Nahrstedt[b], Frank Petereit[b], Barbara Spengler[c], Hilke Winterhoff[a]

extract free of hyperforin and hypericin exerts antidepressant activity

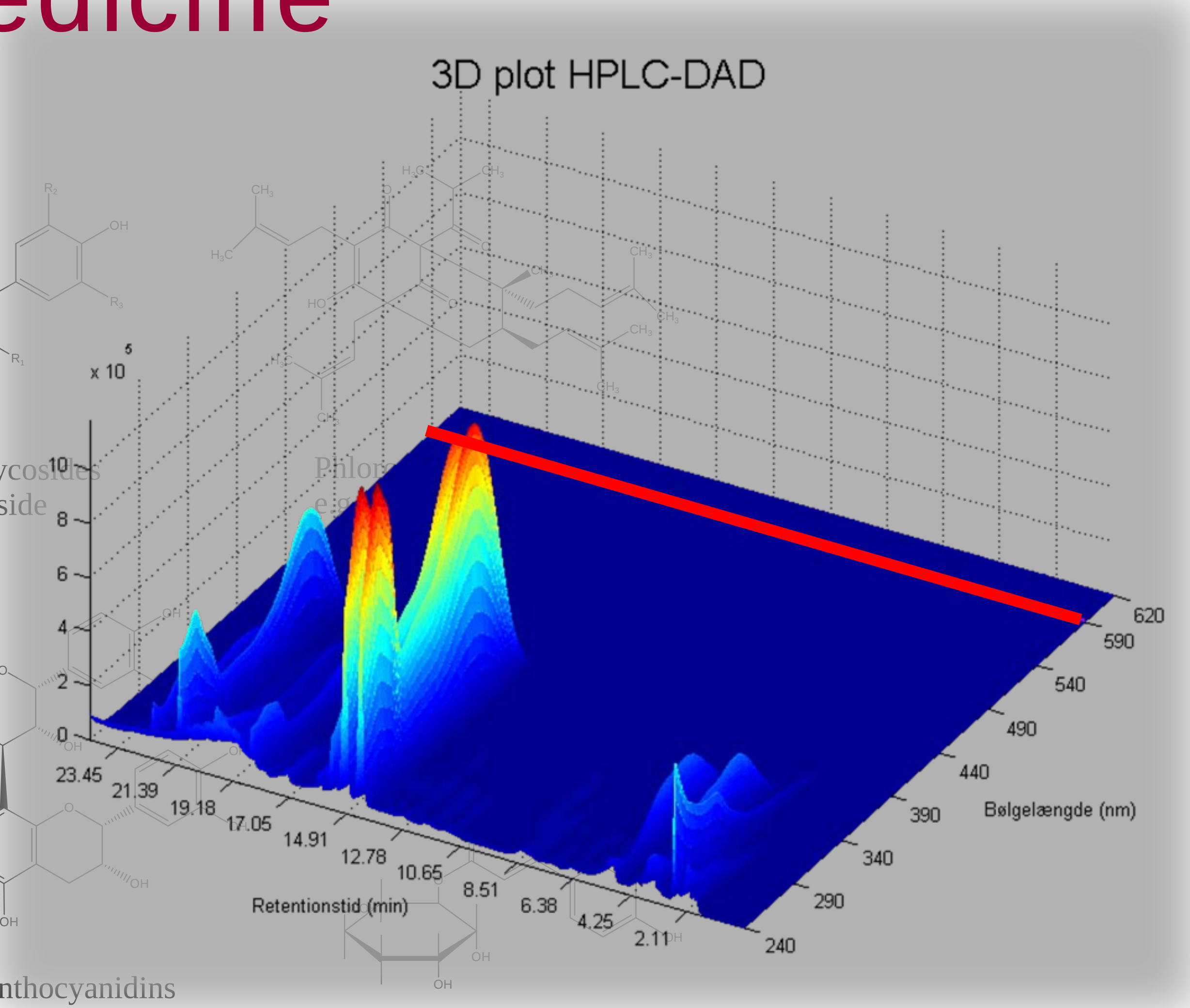
Herbal medicine

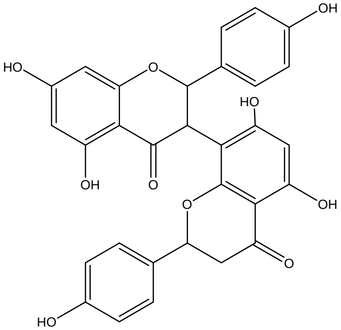


Naphtodianthrones
e.g. hypericin

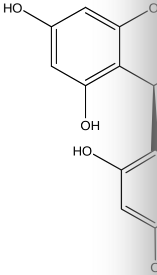


Flavonol glycosides
e.g. hyperoside


3D plot HPLC-DAD



Biflavones
e.g. I3,II8-biapigenin



Protoanthocyanidins
e.g. procyanidin B2

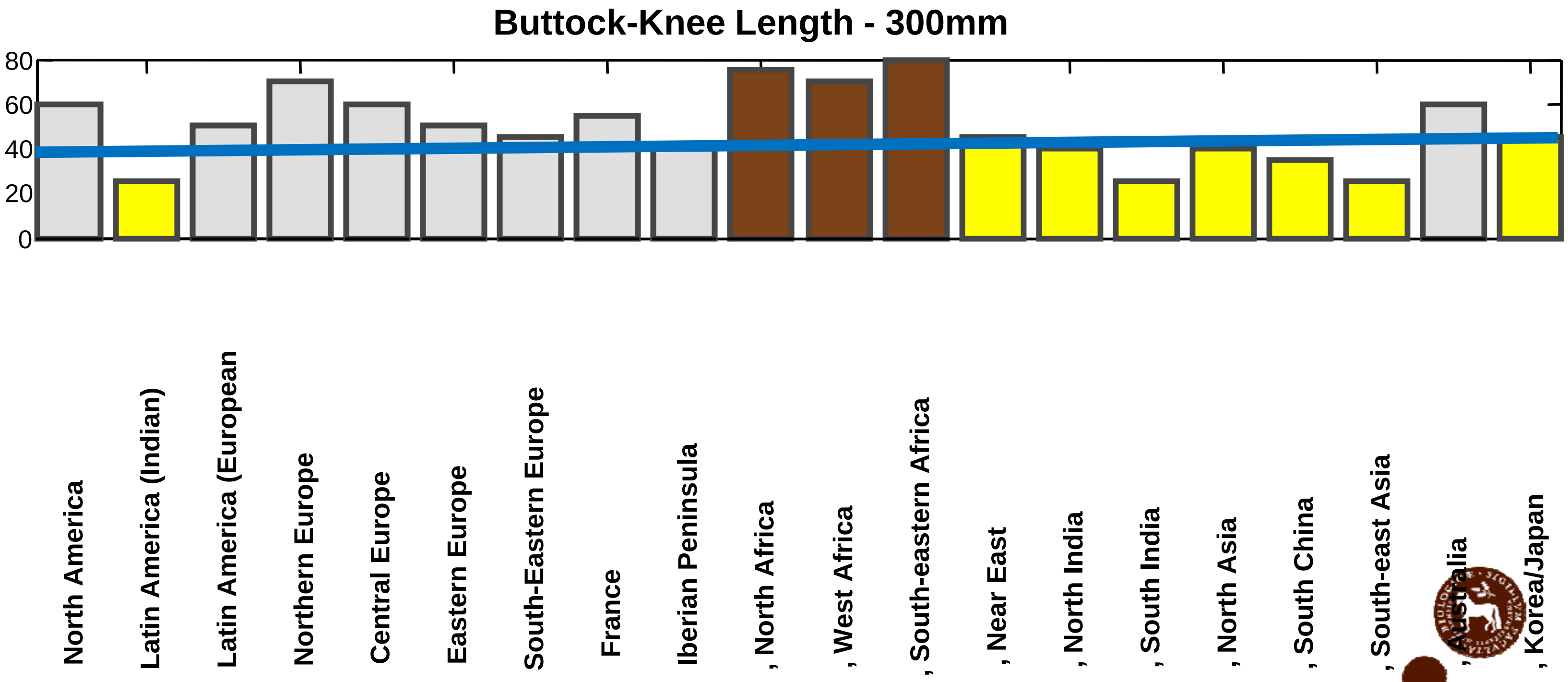
Propylalkans
e.g. chlorogenic acid

Using a single variable is:

Wrong, incorrect, suboptimal, oldfashioned, ..!

Korea overlaps with Caucasian

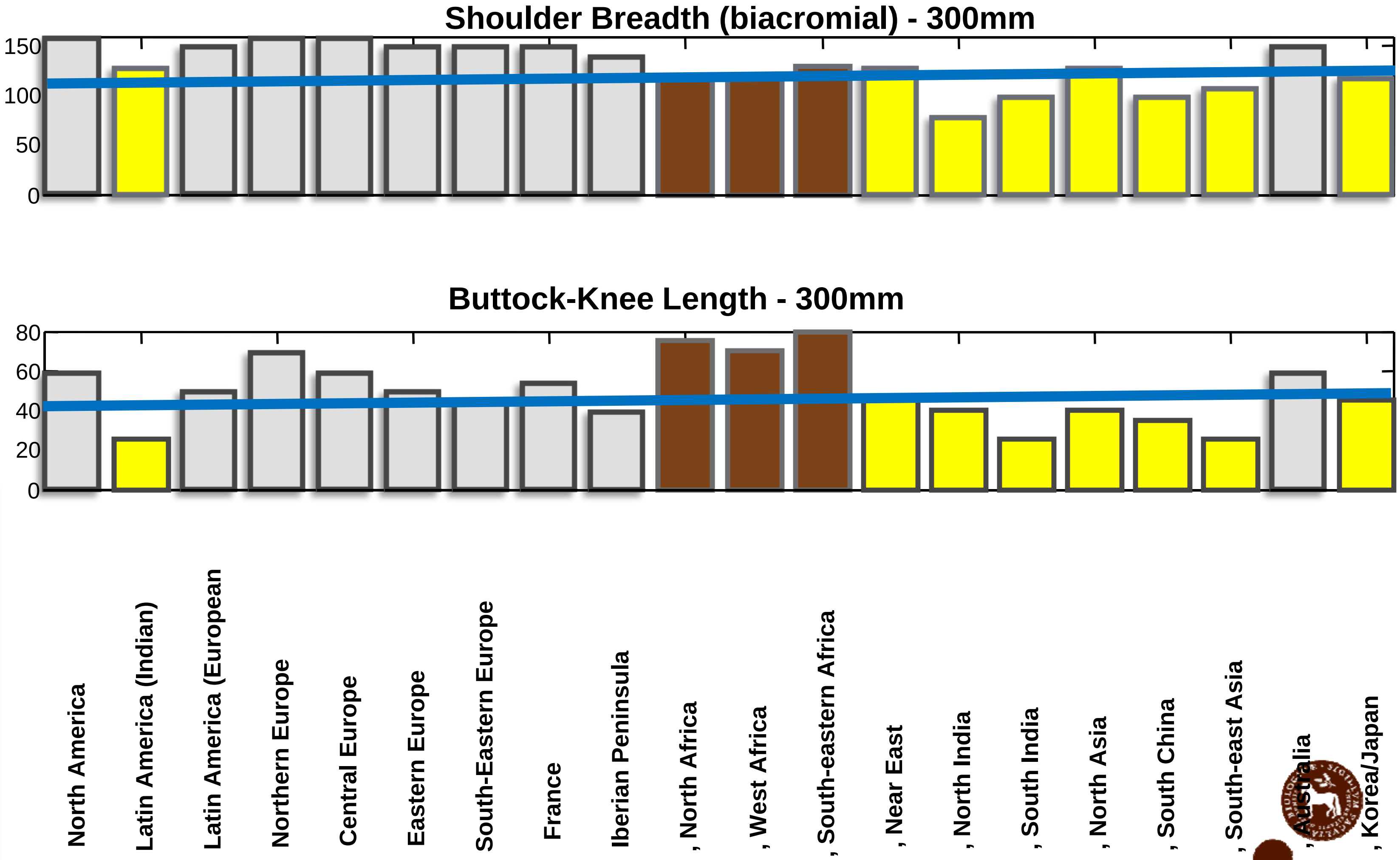
Caucasian
African
Asian

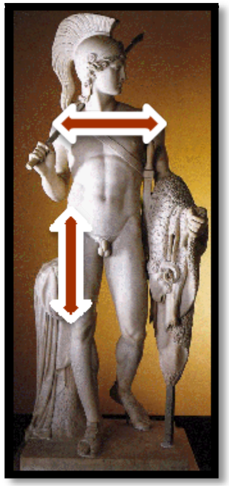


Using a single variable is:

Wrong, incorrect, suboptimal, oldfashioned, ..!

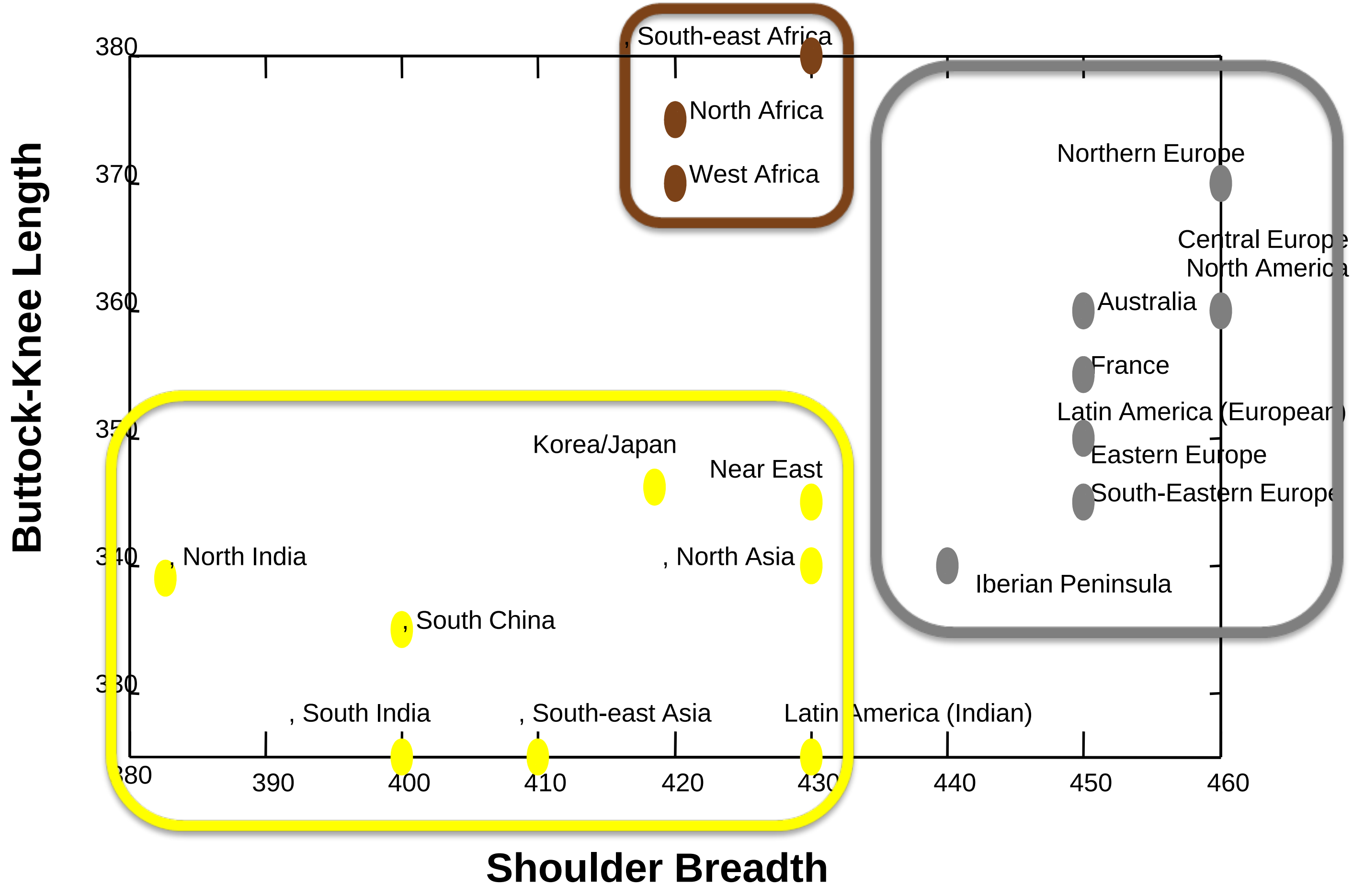
Caucasian
African
Asian





Simply plot the two versus each other

Co-variation = new information that is *not* available in the individual variables

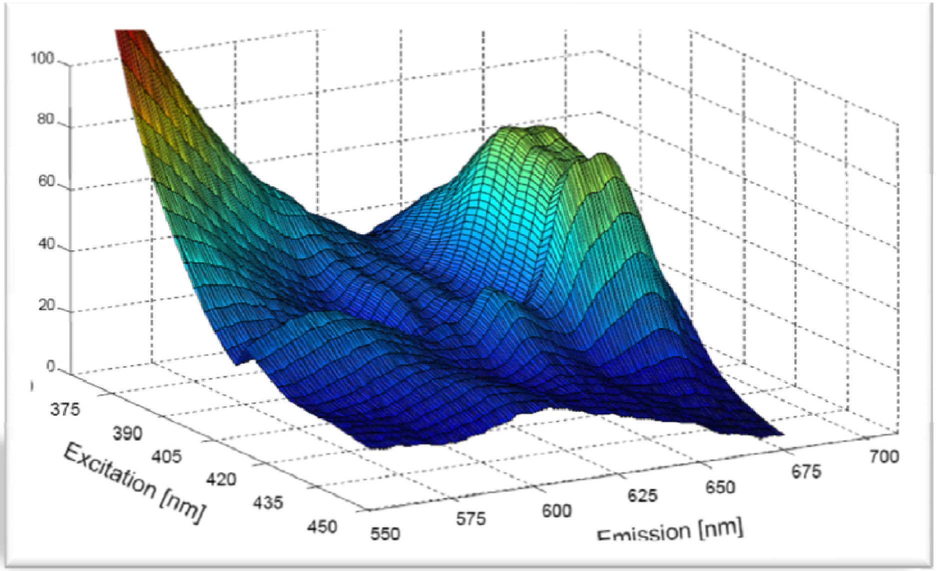


Lee S.,Bro R., Regional Differences in World Human Body Dimensions Theoretical Issues in Ergonomics Science, 2008

Solution:

Don't use univariate methods in complex analysis

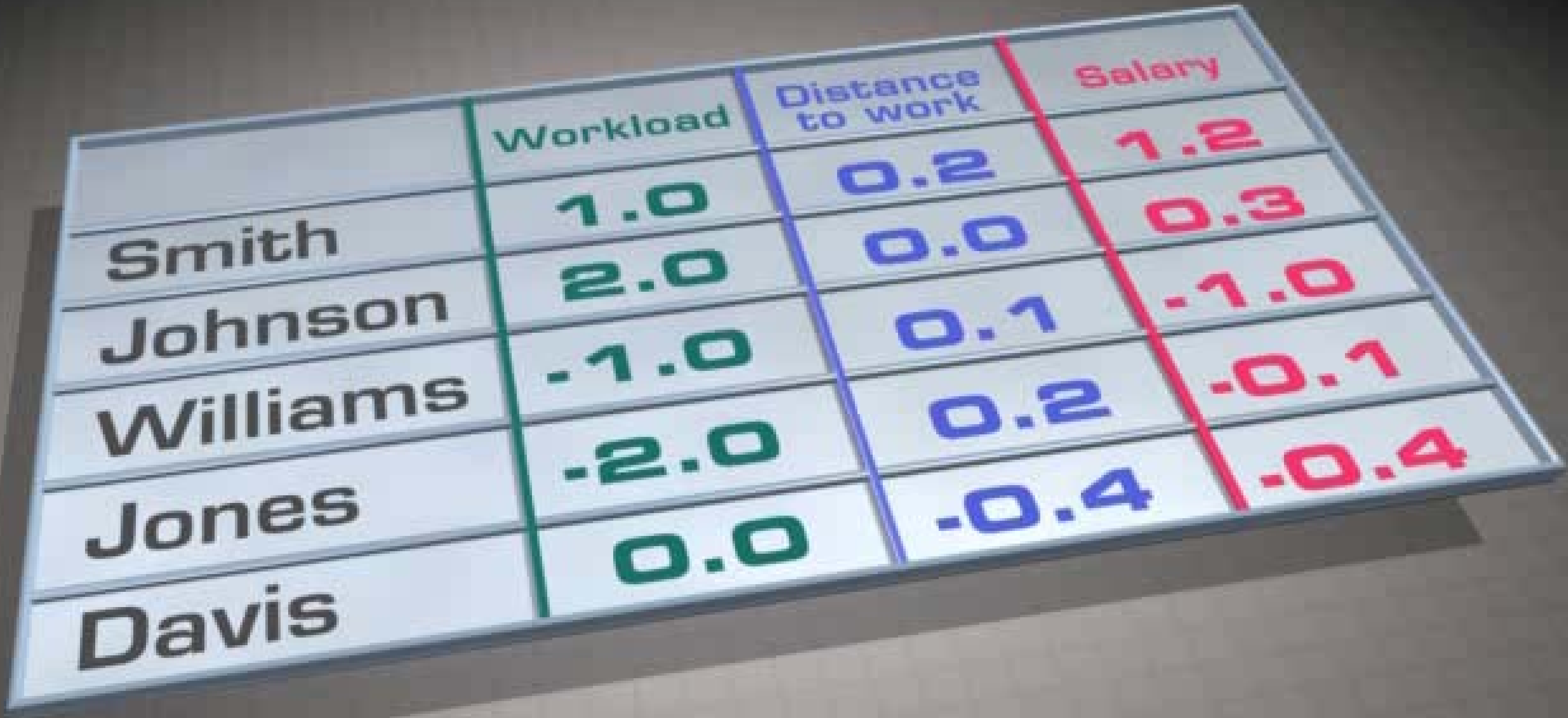
Fluorescence version of a cheese



pH version of a cheese

pH = 6.4

Which would possibly reflect the diversity of cheeses?



	Workload	Distance to work	Salary
Smith	1.0	0.2	1.2
Johnson	2.0	0.0	0.3
Williams	-1.0	0.1	-1.0
Jones	-2.0	0.2	-0.1
Davis	0.0	-0.4	-0.4

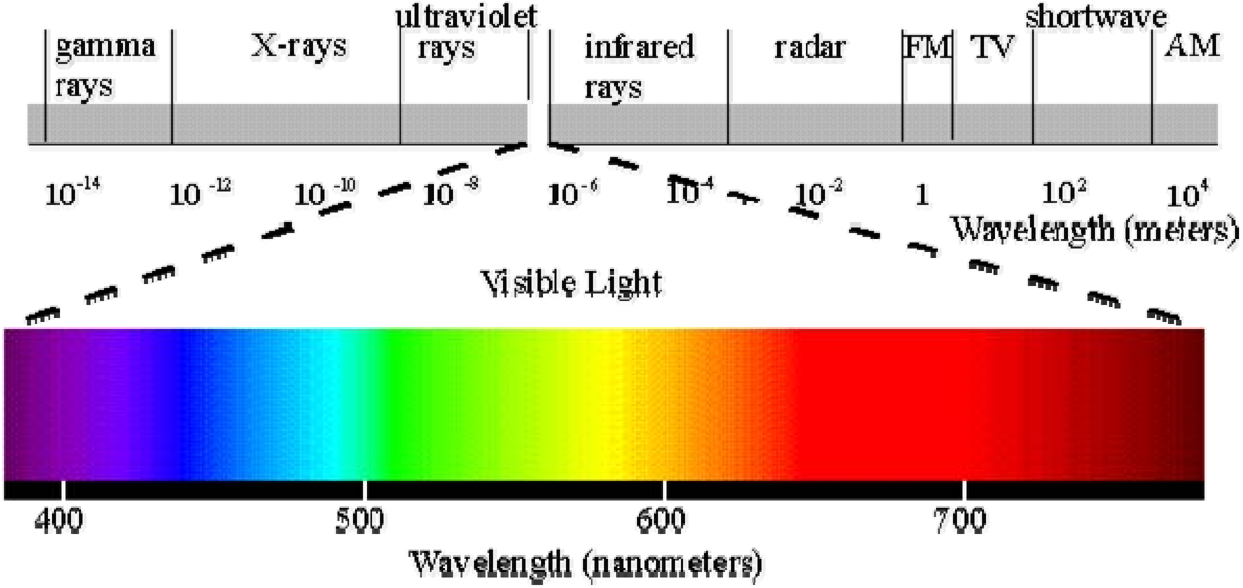
Multivariate data makes it possible to measure directly in blood or in a process

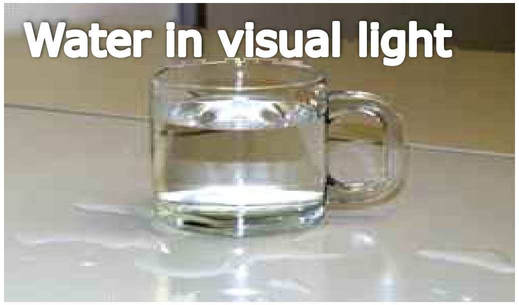
Multivariate data makes it possible to measure directly in blood or in a process

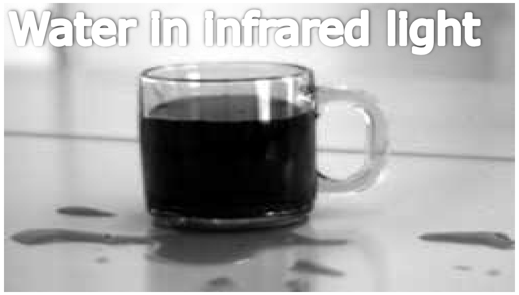
Multivariate data makes it possible to measure *directly* in blood or in a process

Univariate regression needs selective signals

Multivariate data makes it possible to measure *directly* in a complex sample



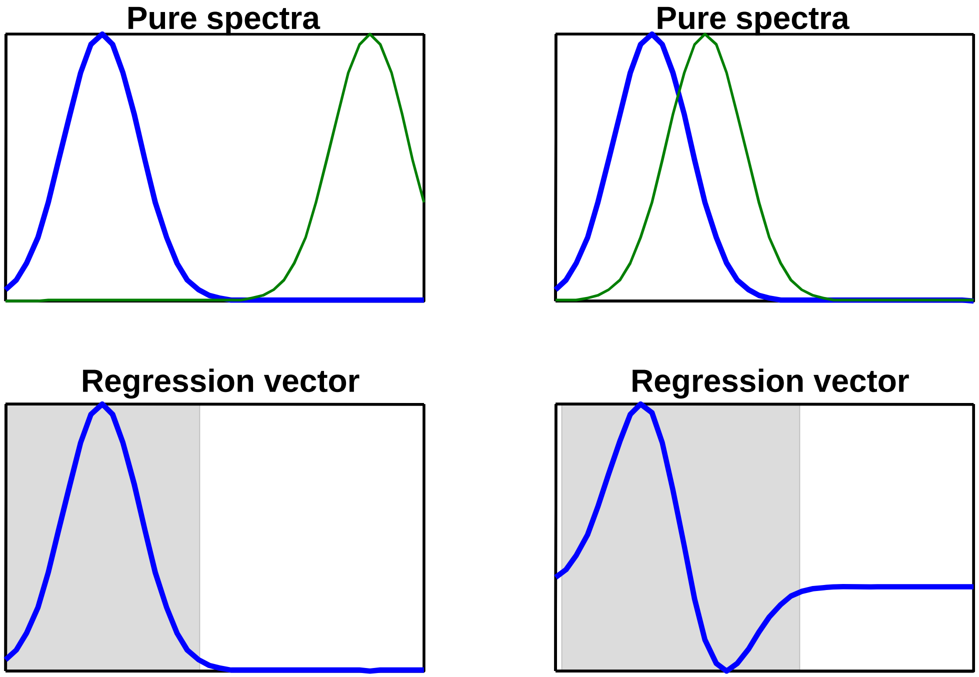




Still problems though!

- Interfering signals (high heels) ok if part of regression model. New ’heels’ not handled

- Regression vector very complicated to interpret



Problems with regression coefs

Signs opposite of physically expected

Even if sign right, indirect and direct (causal) correlations are mixed up

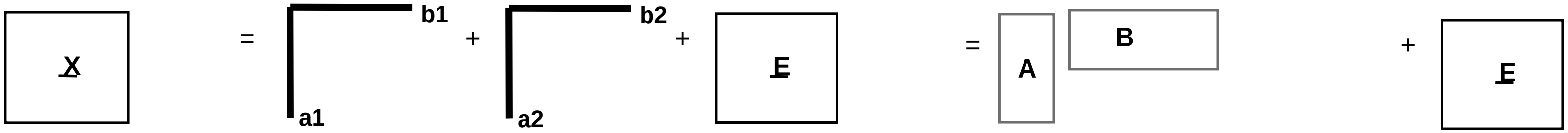
Etc., etc.

Multi-way tensor models

PARallel FACtor analysis

- PCA - bilinear model,

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



Multi-way tensor models

R. A. Harshman. *UCLA working papers in phonetics* 16:1-84, 1970

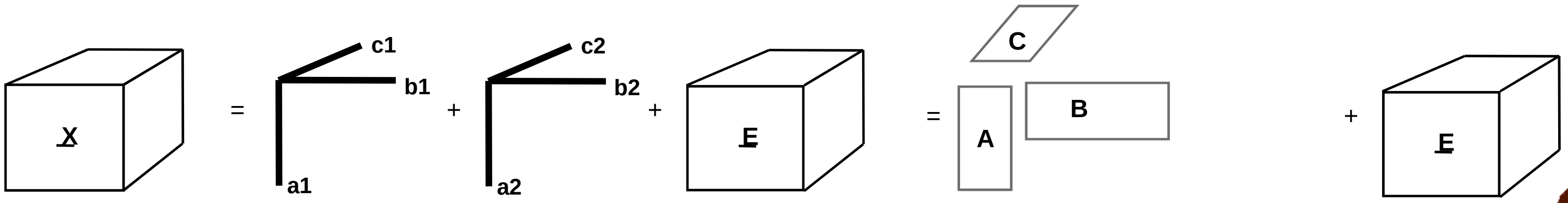
PARallel FACtor analysis

- PCA - bilinear model,

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

- PARAFAC - trilinear model,

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



PARAFAC - algorithm

Given tensor **X** with slabs $\mathbf{X}_k$

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. $diag \mathbf{D}_k = \left\{(\mathbf{B}'\mathbf{B}) * (\mathbf{A}'\mathbf{A})\right\}^{-1} \mathrm{diag}\left(\mathbf{A}'\mathbf{X}_k\mathbf{B}\right), k=1,..,K$

5. Step 2 until relative change in fit is small

$\mathbf{D}_k = \mathrm{diag}(\mathbf{C}(k,:))$

Why ALS?

Simple

Extends to higher order

Handles missing

Handles ML fitting

Constraints:

- Nonnegativity
- Unimodality
- Orthogonality
- Linear constraints
- Fixed parameters
- Smoothness
- Functional
- etc

PARAFAC - uniqueness

- **No rotational freedom as in PCA**

  If the measured data follows a PARAFAC model, PARAFAC can retrieve the underlying parameters – i.e. solve the cocktail party effect/inverse problem.

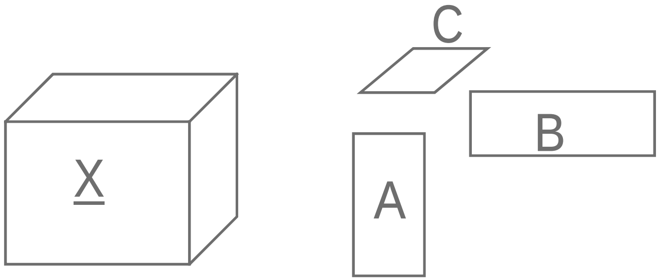
Means no outliers

- **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 **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.

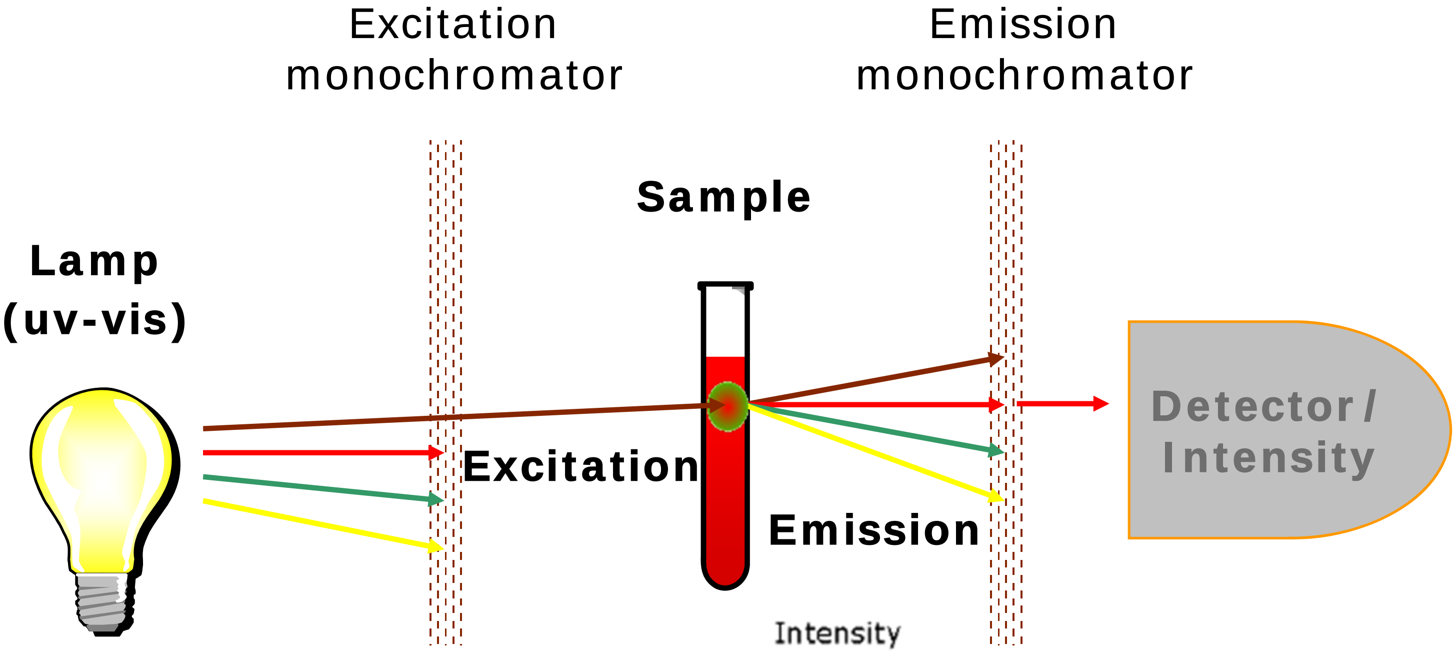
PARAFAC - uniqueness

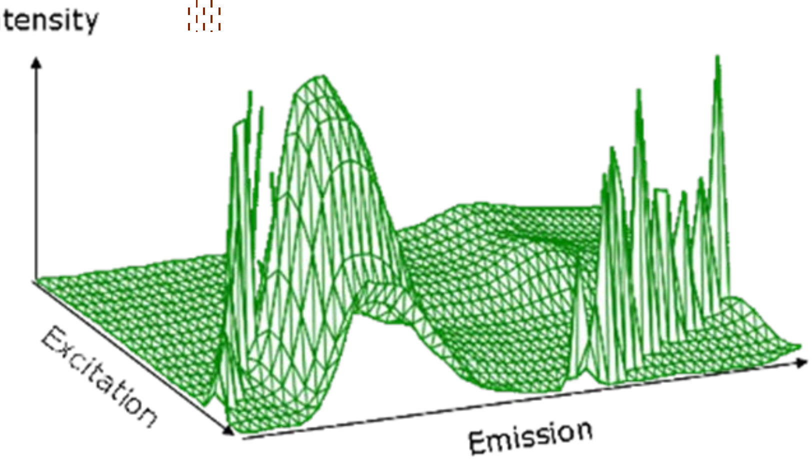
Uniqueness – funny stuff

For example, an 8-component PARAFAC model of a 6×6×6 array is unique

- I.e. six observations – eight different components!
- This compares to getting 8 PCA components from a 6×36 matrix!

Fluorescence spectroscopy

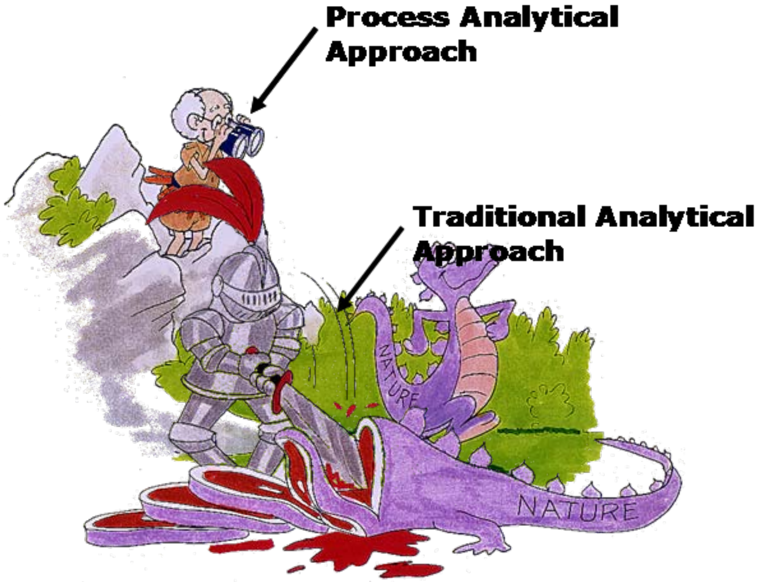




Excitation-emission matrix – a chemical fingerprint

Online fermentation monitoring

- Quality (enzyme) measured rarely =>
- Quality control not possible



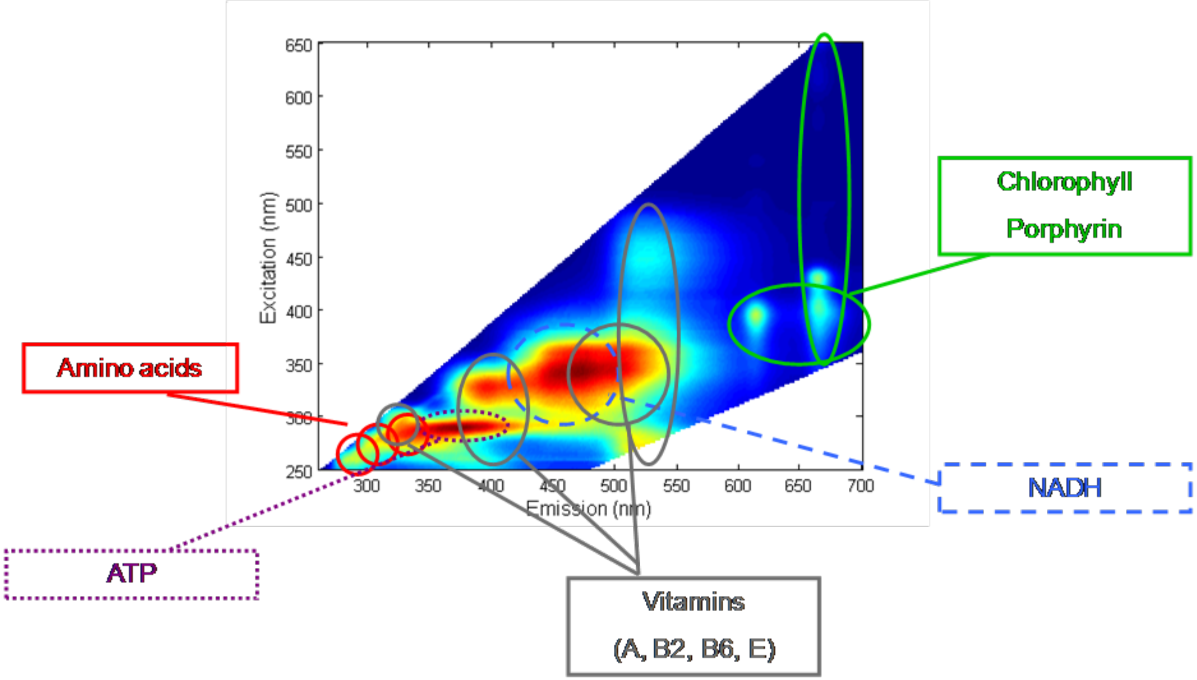
Process Analytical Approach

Traditional Analytical Approach

Remote
Non-destructive
Exploratory
Fast
Multivariate
Multi-relevant

Invasive
Destructive
Deductive
Time-consuming
Univariate

Online fermentation monitoring



Amino acids

ATP

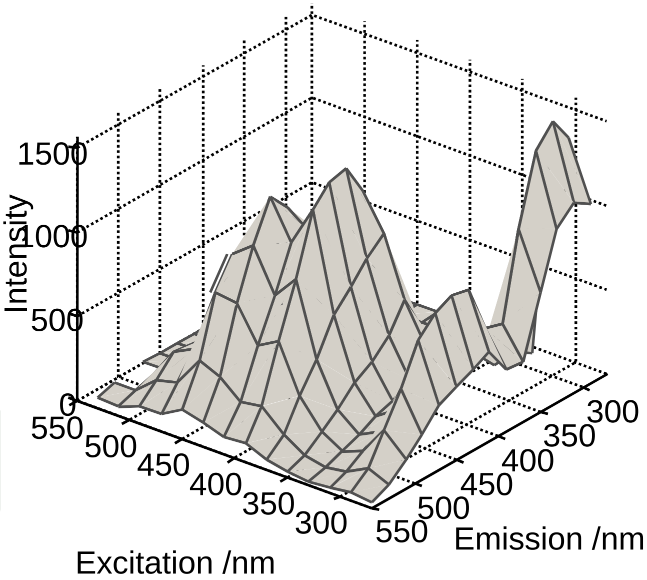
Vitamins

(A, B2, B6, E)

NADH

Chlorophyll

Porphyrin



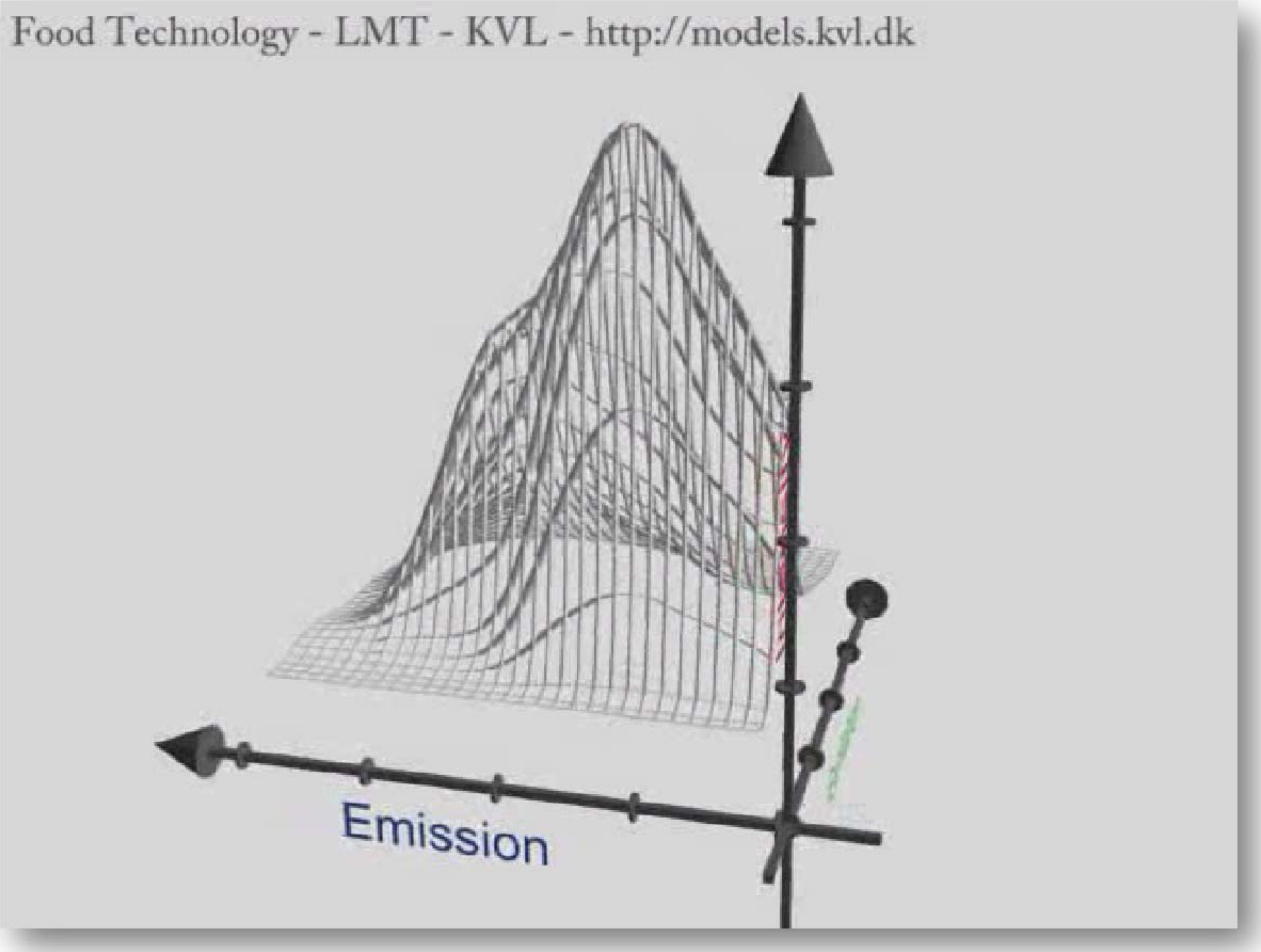
Chemometrics and Intelligent Laboratory Systems 84 (2006) 106–113

Real-time monitoring and chemical profiling of a cultivation process

Peter P. Mortensen [a,b,*], Rasmus Bro [c]

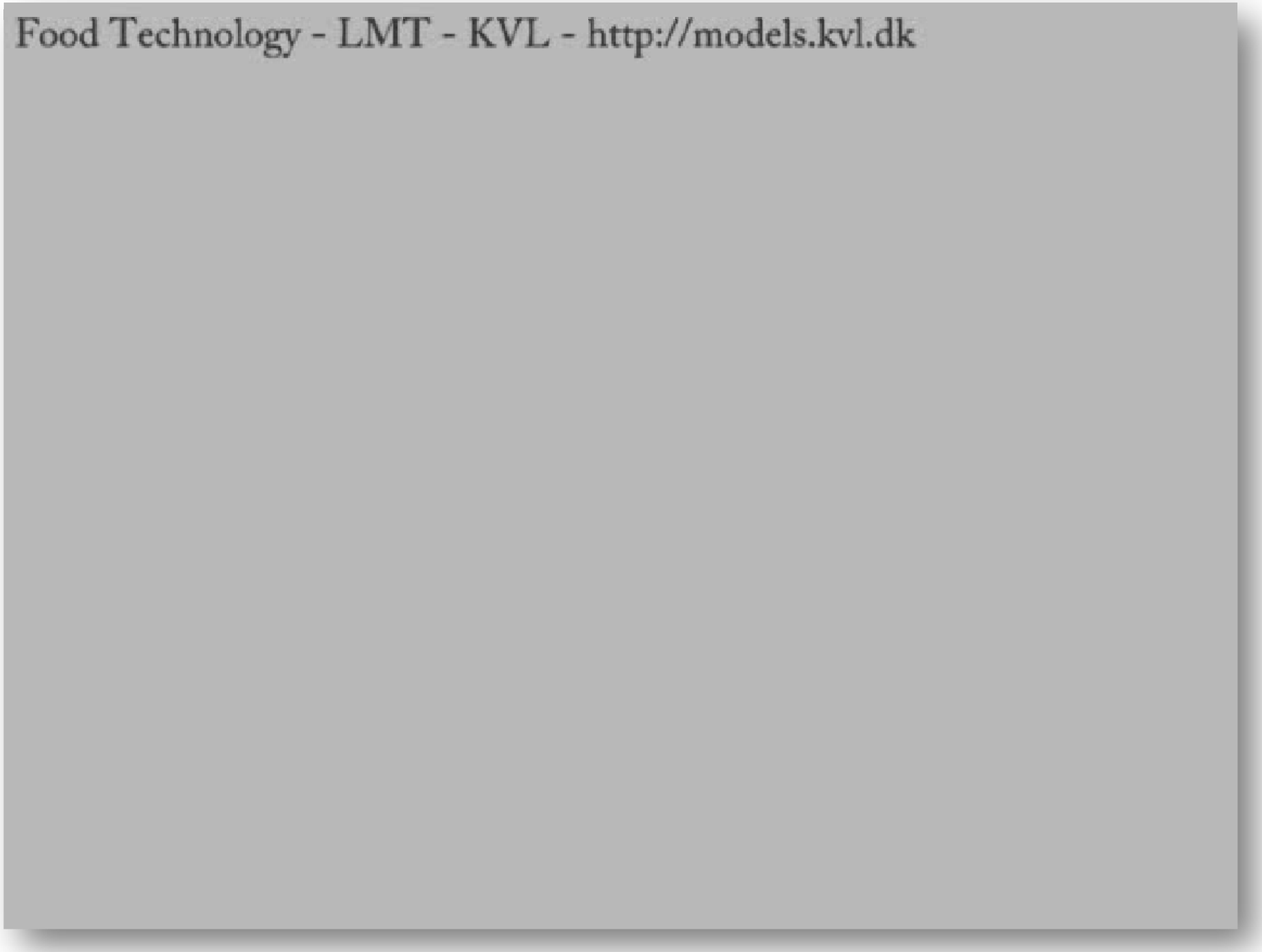
Online fermentation

Mathematical Chromatography



Online fermentation

Mathematical Chromatography



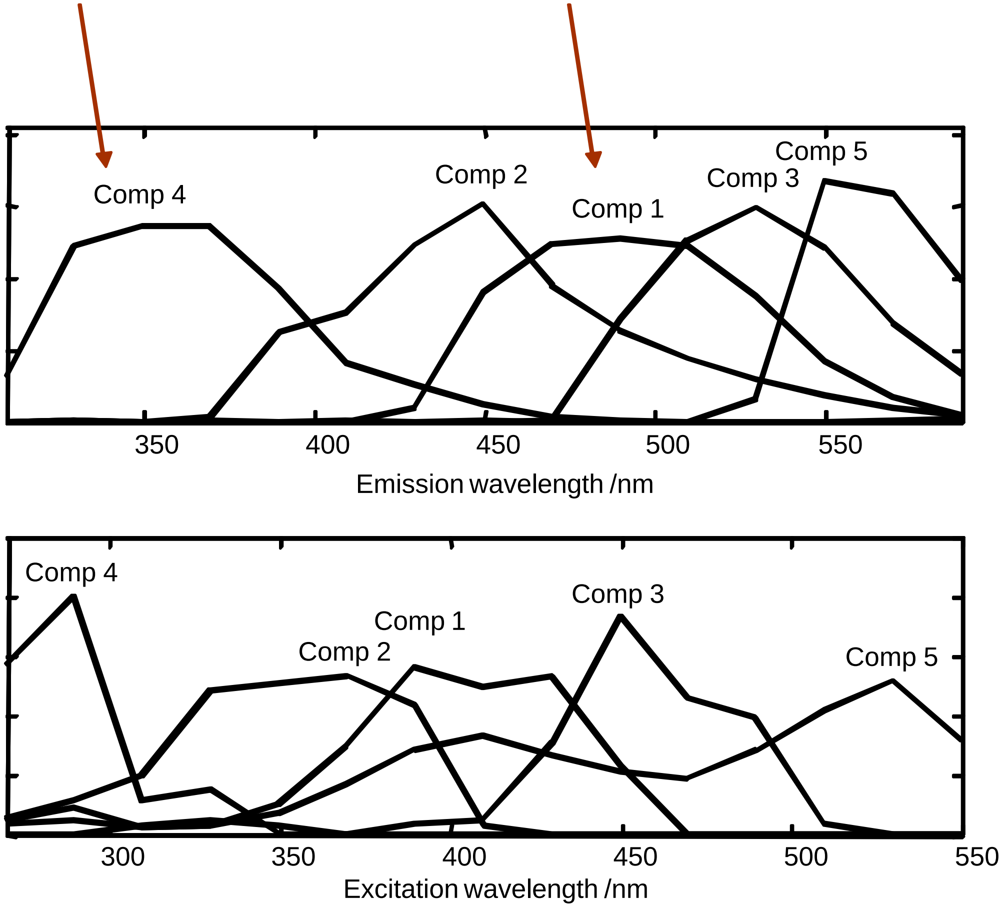
Food Technology - LMT - KVL - http://models.kvl.dk

Online fermentation

Mathematical Chromatography

Protein

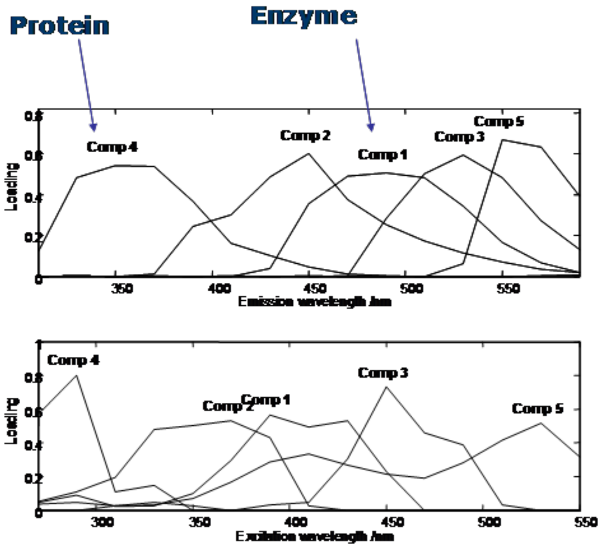
Enzyme

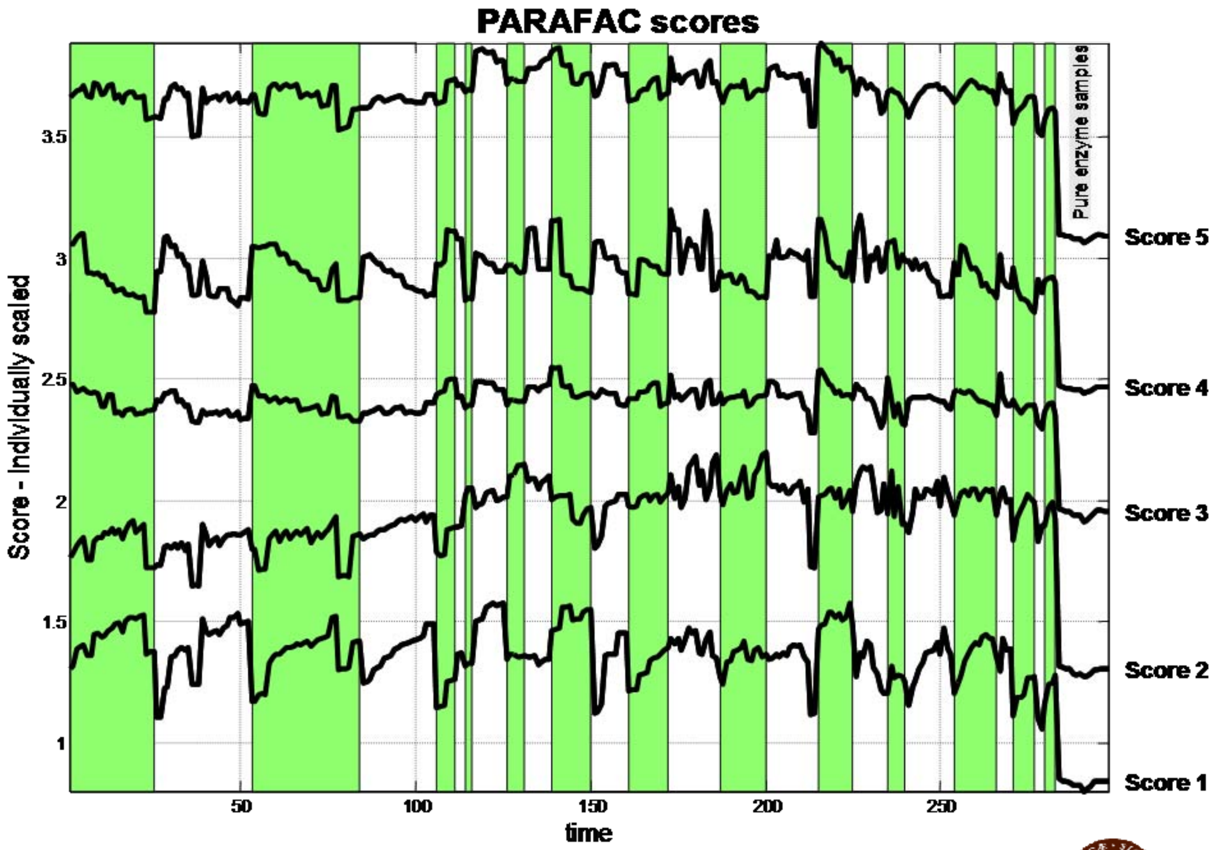


Critical *process and quality parameters* can be *directly* identified

Online fermentation
Mathematical Chromatography

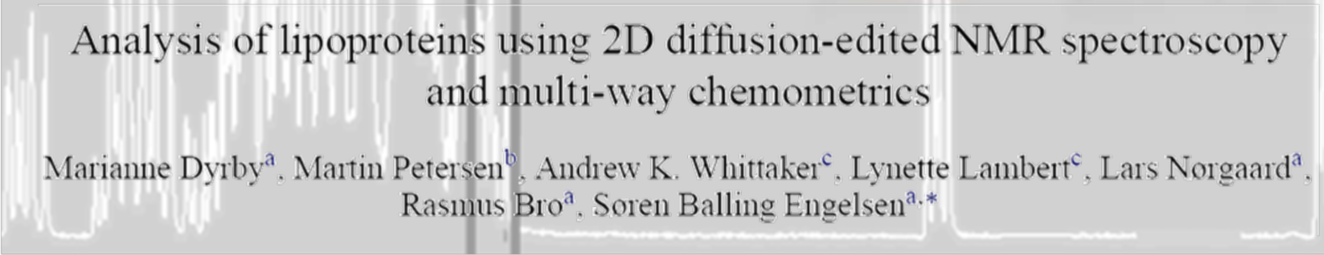
Process monitored frequently





Using PARAFAC for high resolution NMR

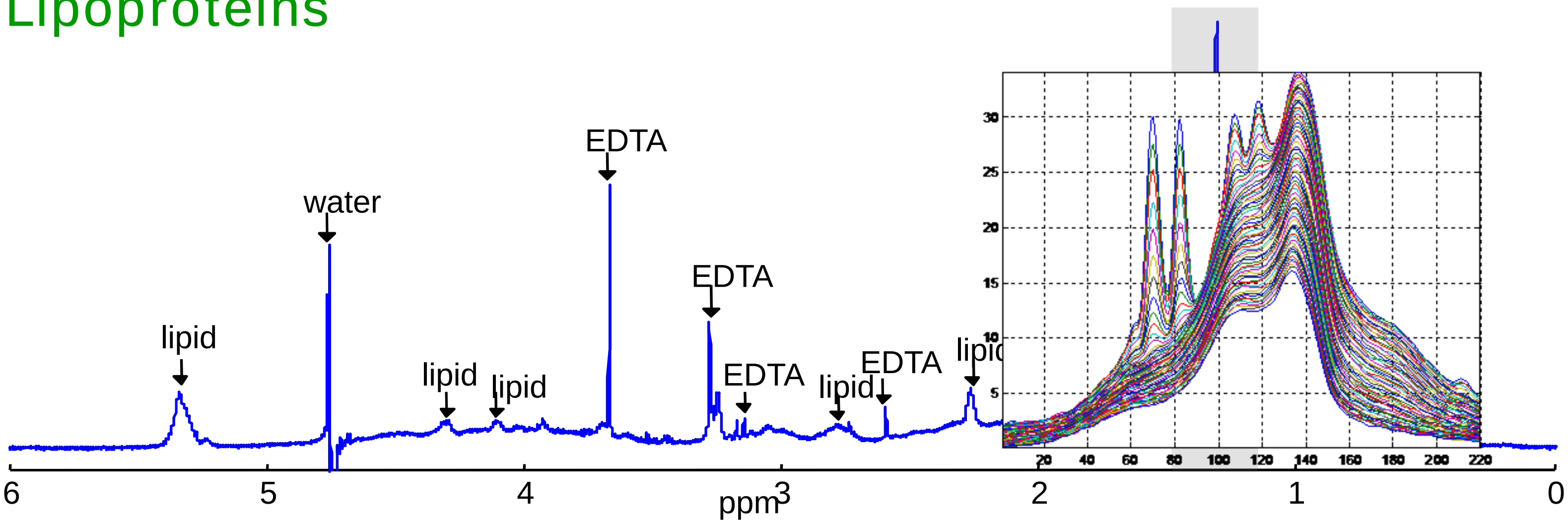
Fast and direct lipoprotein profiling

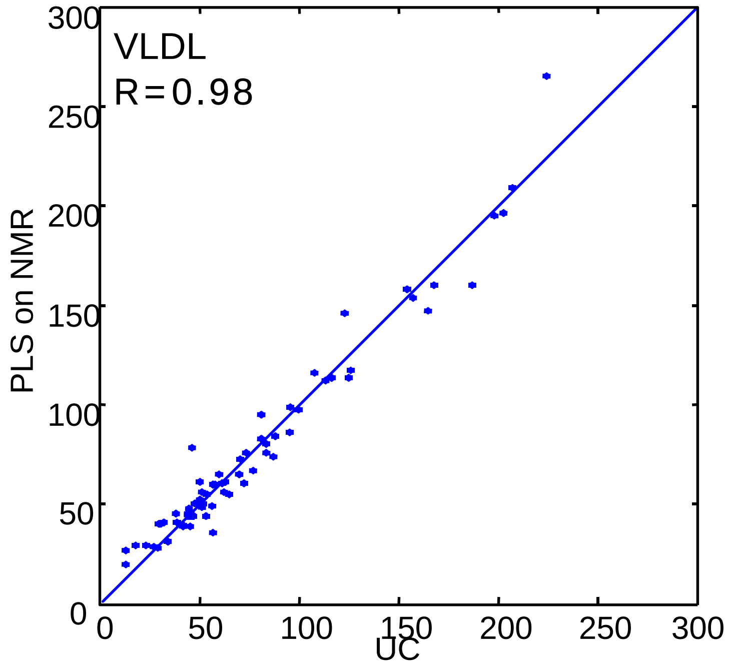


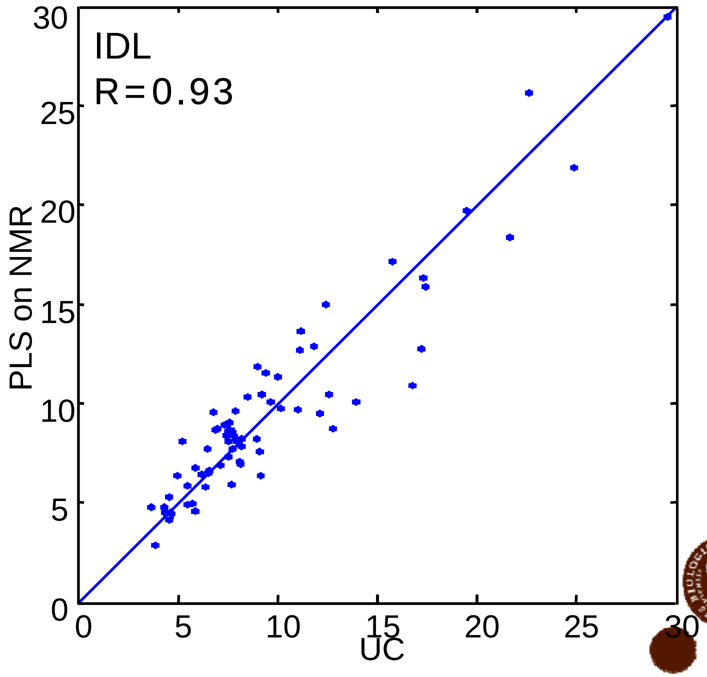
Analysis of lipoproteins using 2D diffusion-edited NMR spectroscopy and multi-way chemometrics

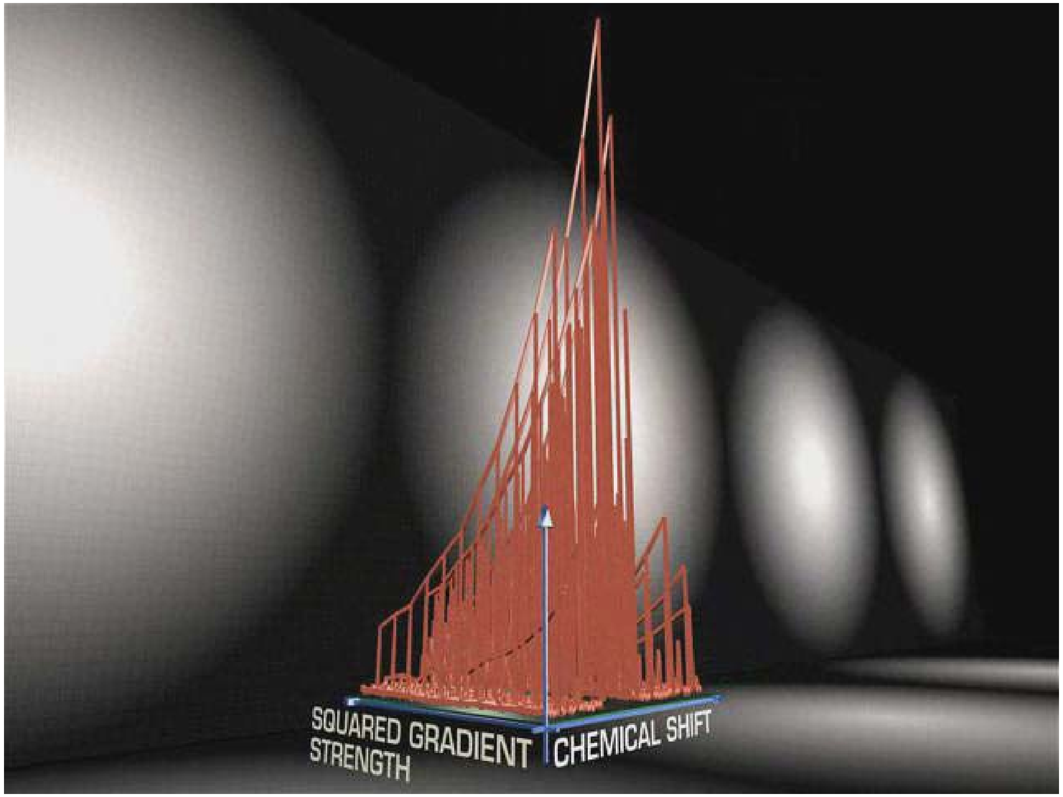
Marianne Dyrby[a], Martin Petersen[b], Andrew K. Whittaker[c], Lynette Lambert[c], Lars Nørgaard[a],
Rasmus Bro[a], Søren Balling Engelsen[a,*]

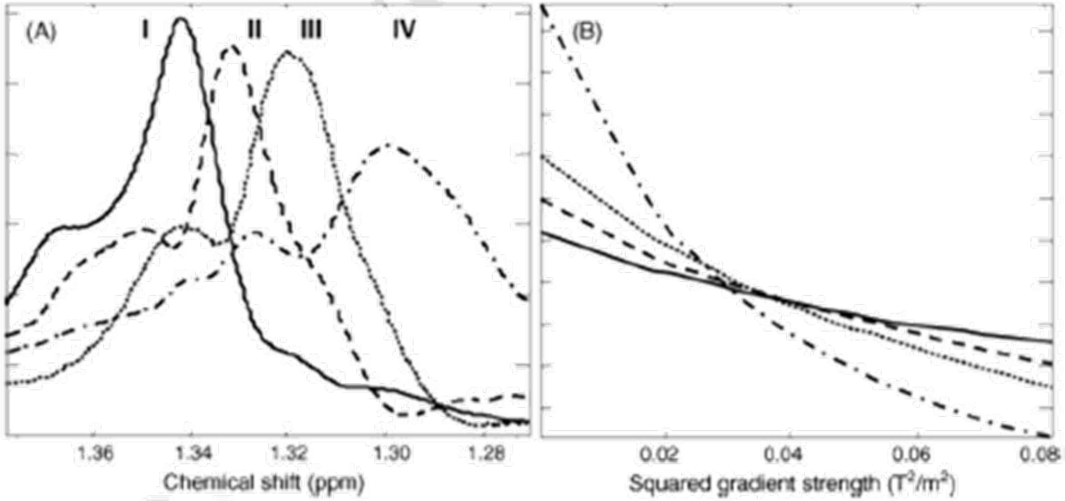
Lipoproteins











Understanding why

PARAFAC is mathematical chromatography

Mathematical chromatography eliminates major problems in multivariate analysis:

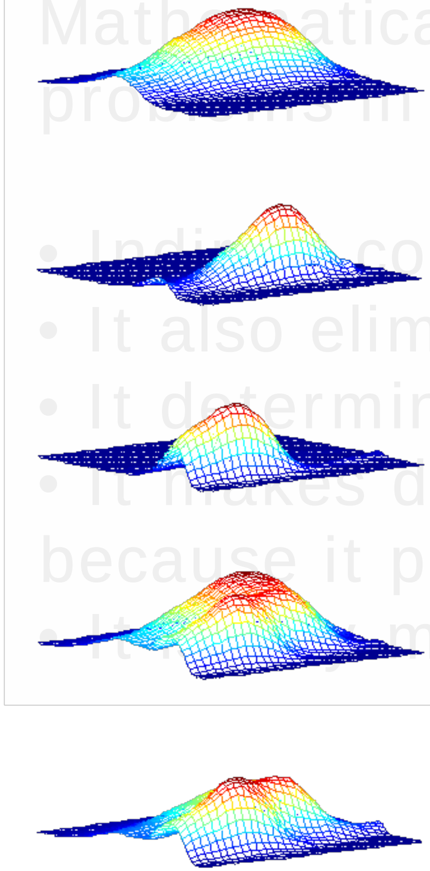
- Indirect correlations stemming from rotational freedom
- It also eliminates outliers
- It determines underlying sources
- Simpler because it provides a chemical model
- It is way more noise insensitive

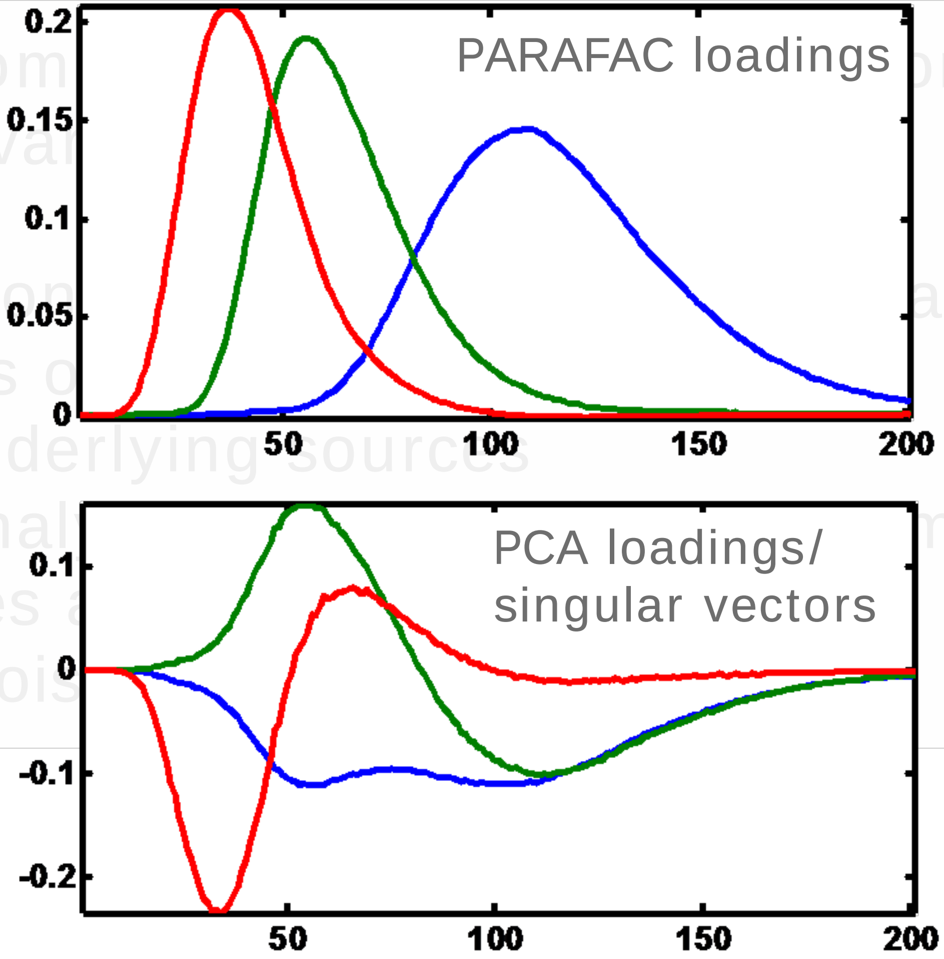
PARAFAC is mathematical chromatography

Mathematical chrom or

problems in multivar

- Indirect correlation al freedom
- It also eliminates o
- It determines underlying sources
- It makes data anal mists because it provides a
- It more nois

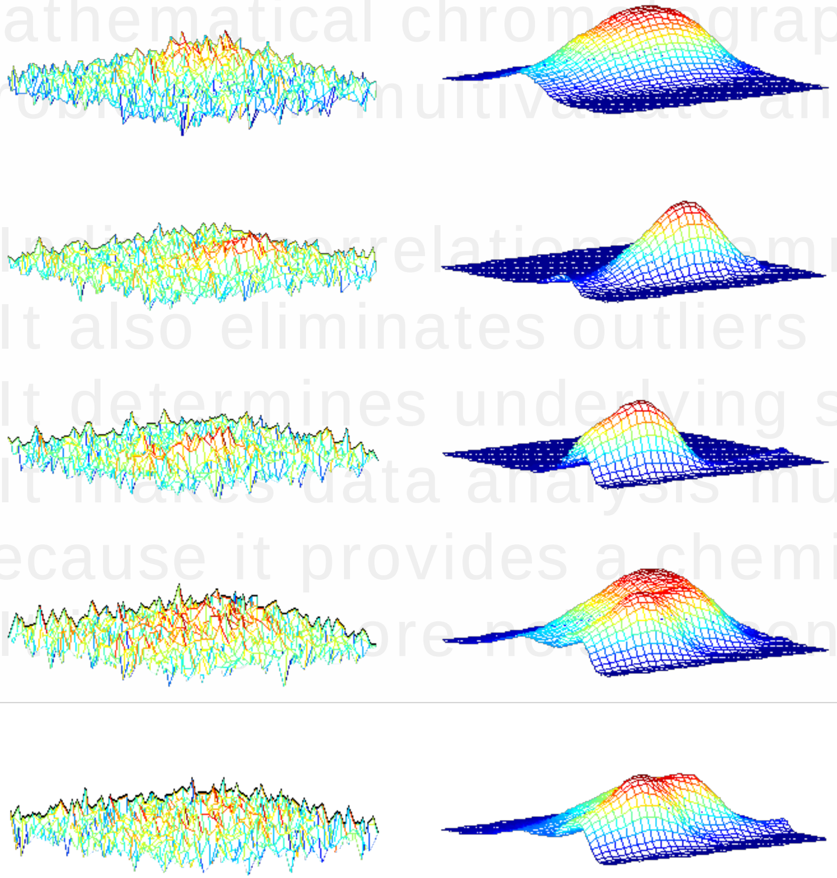




PARAFAC is mathematical chromatography

Mathematical chromatography eliminates major problems in multivariate analysis:

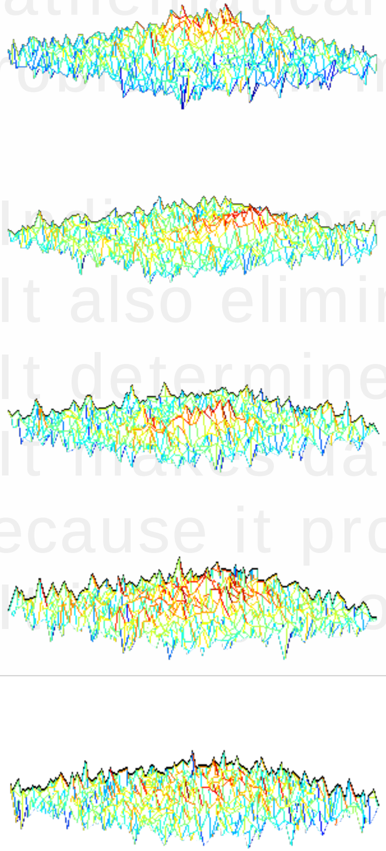
- Indirect correlations stemming from rotational freedom
- It also eliminates outliers
- It determines underlying sources
- It makes data analysis much simpler for chemists because it provides a chemical model
- It is more noise insensitive

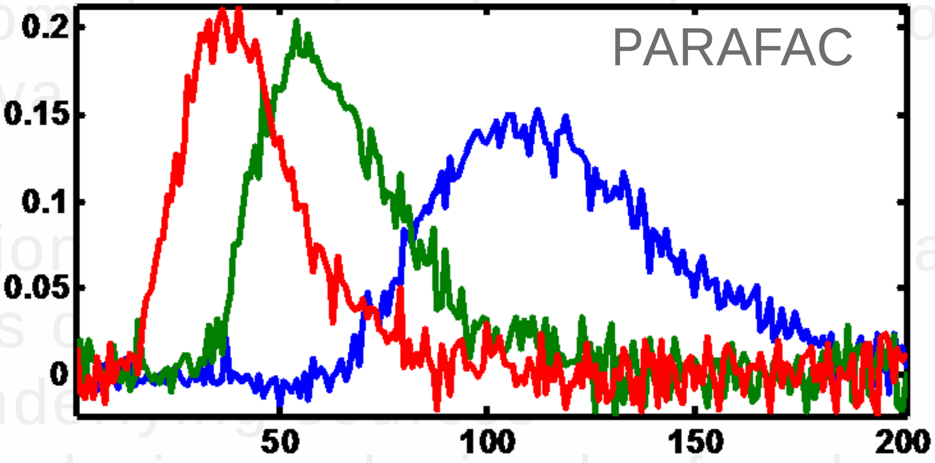


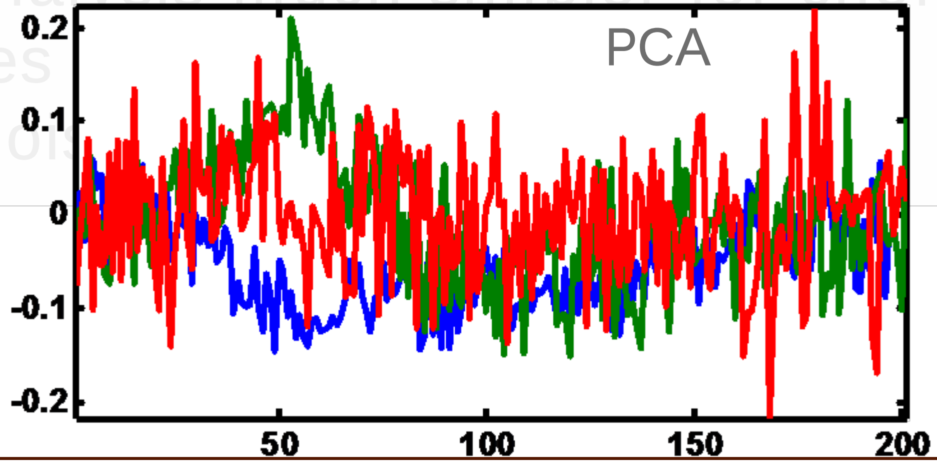
PARAFAC is mathematical chromatography

Mathematical chrom... ...or
prob... ...multiva...

- I... ...rrelation... ...al freedom
- It also eliminates c...
- It determines unde... ...lying sou...
- It makes data analysis much simpler for chemists because it provides
- I... ...ore nois...

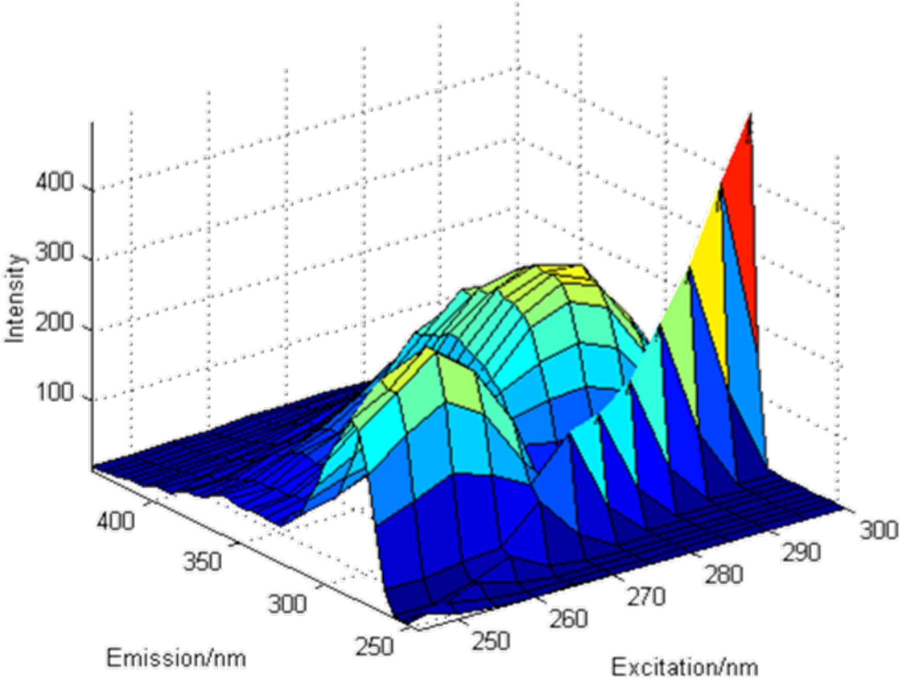






Data analysis requires good data – g.i.g.o.

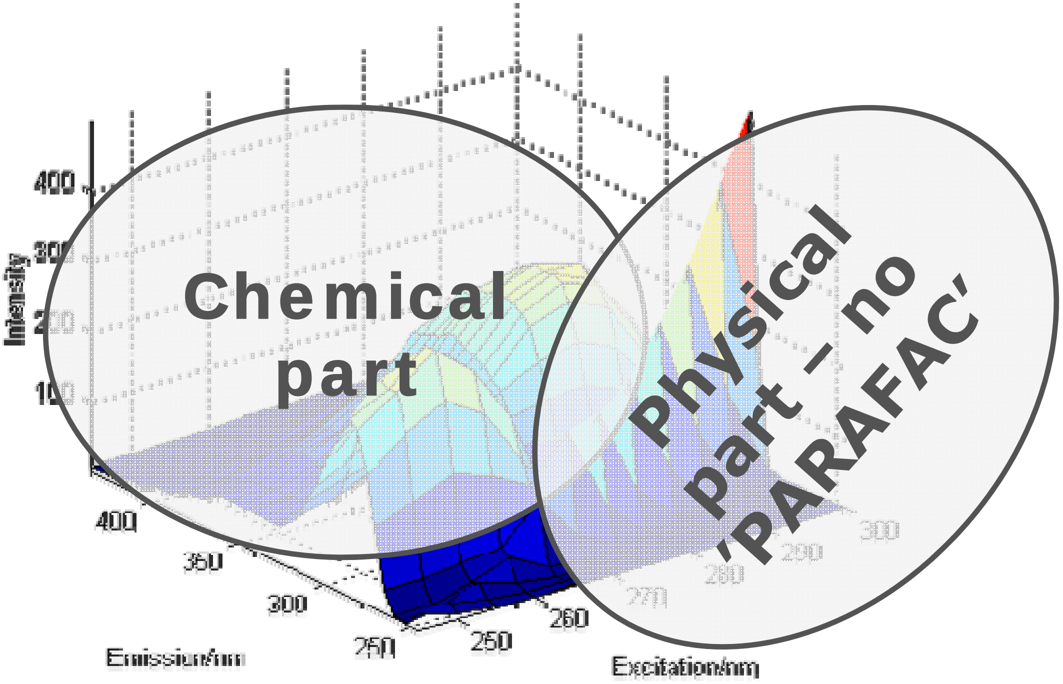
Data analysis requires good data – g.i.g.o.



Example

Fluorescence excitation-emission matrix contains chemical information that PARAFAC can handle and physical scattering signals that do not fit PARAFAC

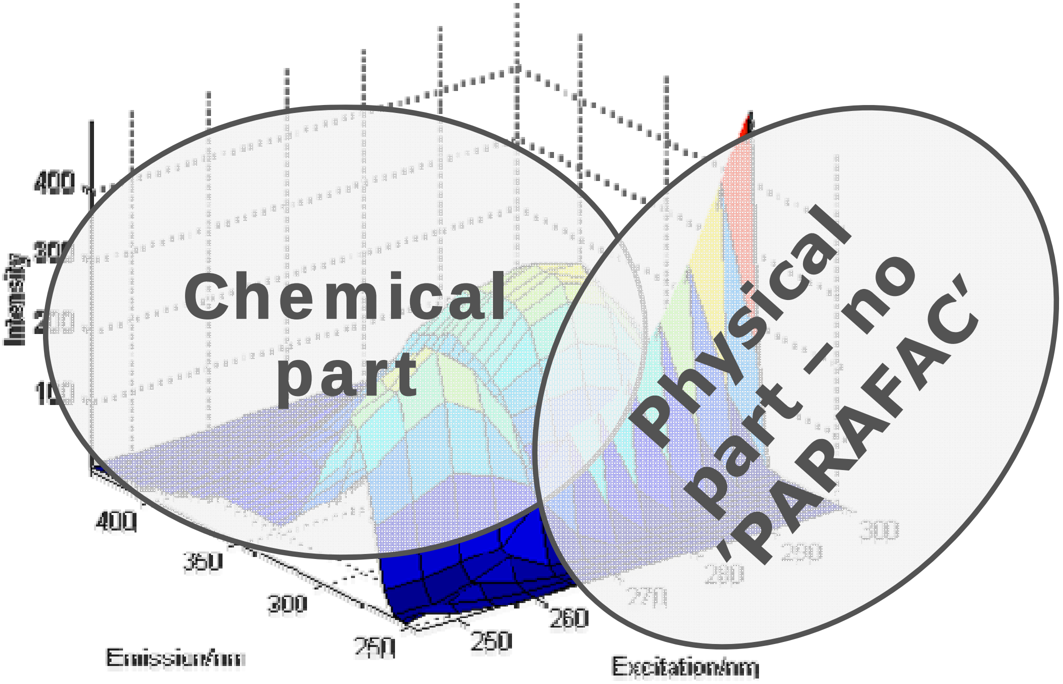
Knowing your data



Chemical part

Physical part – no 'PARAFAC'

Knowing your data



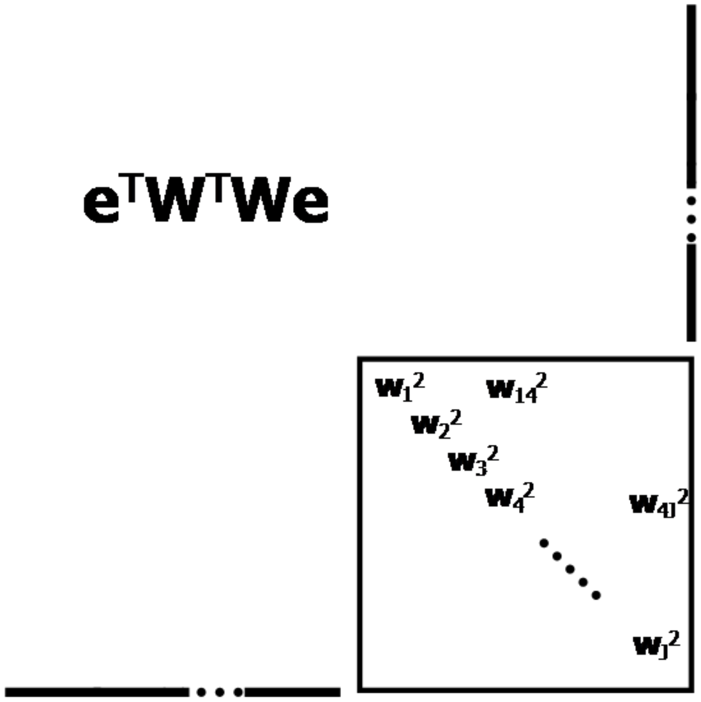
Chemical part

Physical part – no 'PARAFAC'

Knowing your data

MILES – maximum likelihood

A way to downweigh areas of less importance by extending least squares fit to weighted and off-diagonal-weighted least squares



MILES – maximum likelihood

- Algorithm **MILES** (Maximum likelihood via Iterative Least squares EStimation) based on Majorization
- Enables weighted least squares and maximum likelihood fitting of any model which has a least squares algorithm

Given vectorized data **x** and weights **W**

1. Initialize model, $\mathbf{m}_0$, with LS, set c := 0;
2. $\mathbf{q} = \mathbf{m}_c + 1/\beta\ \mathbf{W}^{\mathrm{T}}\mathbf{W}(\mathbf{x} - \mathbf{m}_c)$
3. $\mathbf{m}_{c+1} = \underset{\mathbf{m}\in\Upsilon}{\operatorname{argmin}} \left\| \mathbf{m} - \mathbf{q} \right\|_F^2$
4. c := c+1; go to step 2 until convergence

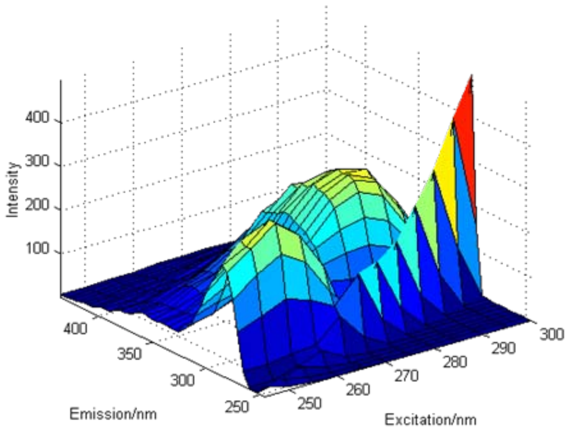
Calculate **q**

Fit LS model to **q** instead of to data

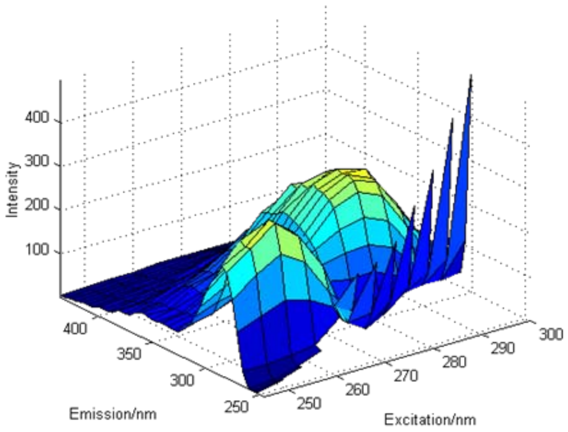
- 21 samples containing L-phenylalanine, L-3,4-dihydroxy-phenyl-alanine (DOPA), 1,4-dihydroxy-benzene & L-tryptophan

- Three types of unwanted variation
  - Measurement error (~iid Gaussian)
  - Rayleigh and Raman scatter
  - Non-chemical area

Typical sample



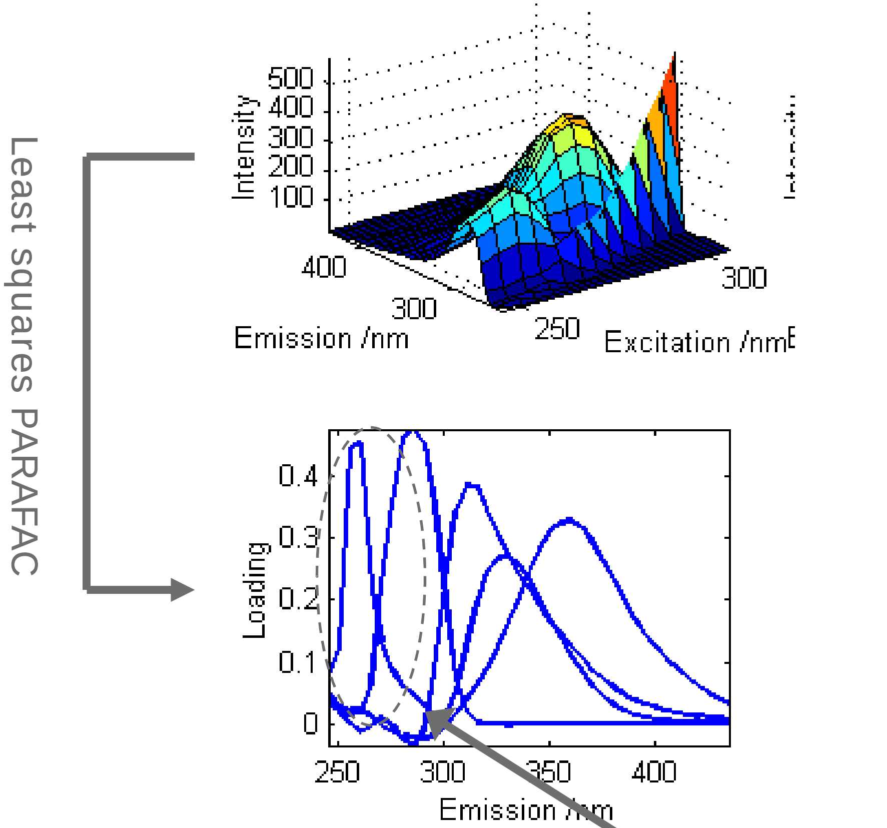
Typical sample (emi<exc removed)



Baunsgaard D, Factors affecting 3-way modelling (PARAFAC) of fluorescence landscapes, The Royal Veterinary & Agricultural University, 1999

PARAFAC results

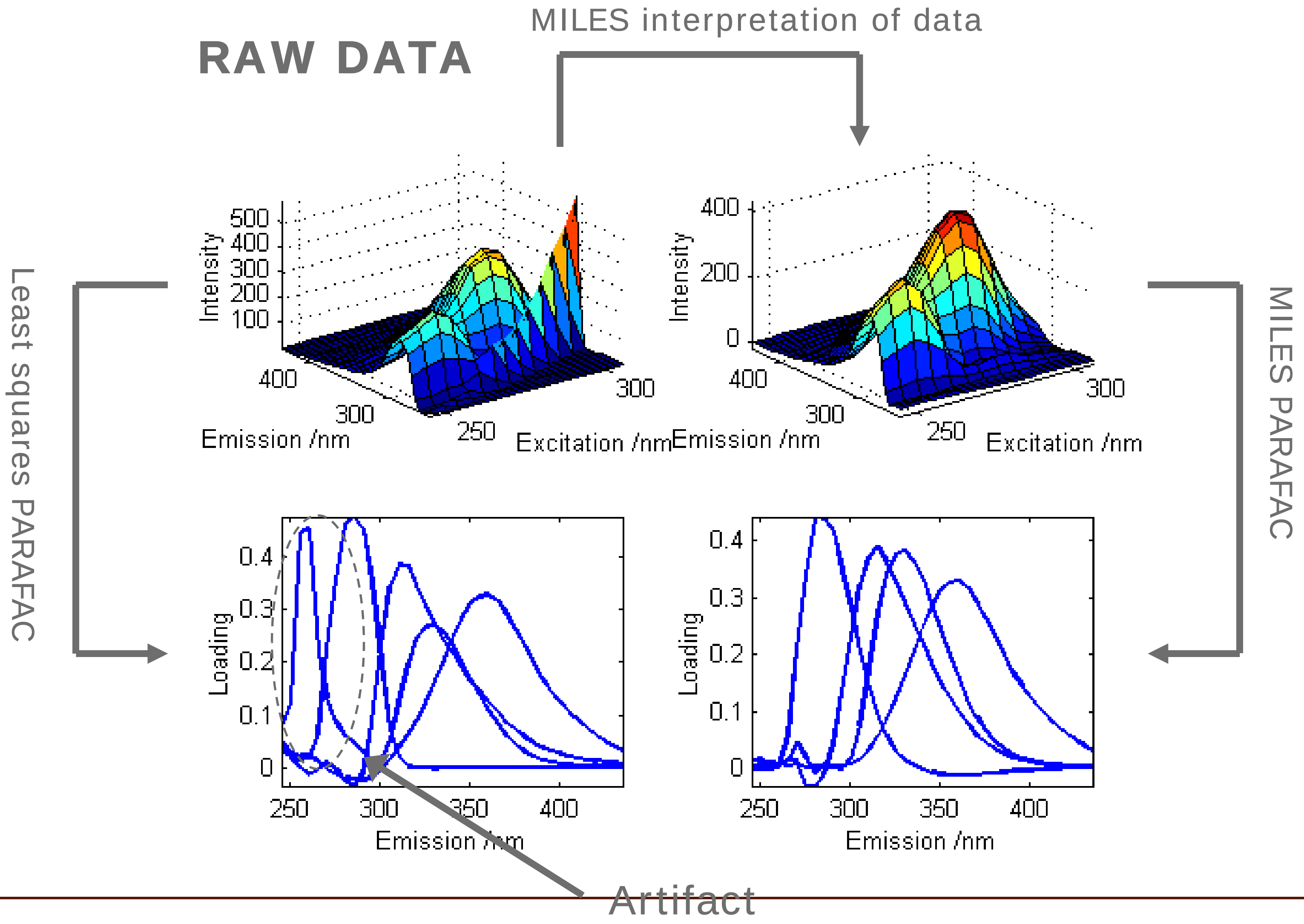
RAW DATA



Least squares PARAFAC

Artifact

PARAFAC results



RAW DATA

MILES interpretation of data

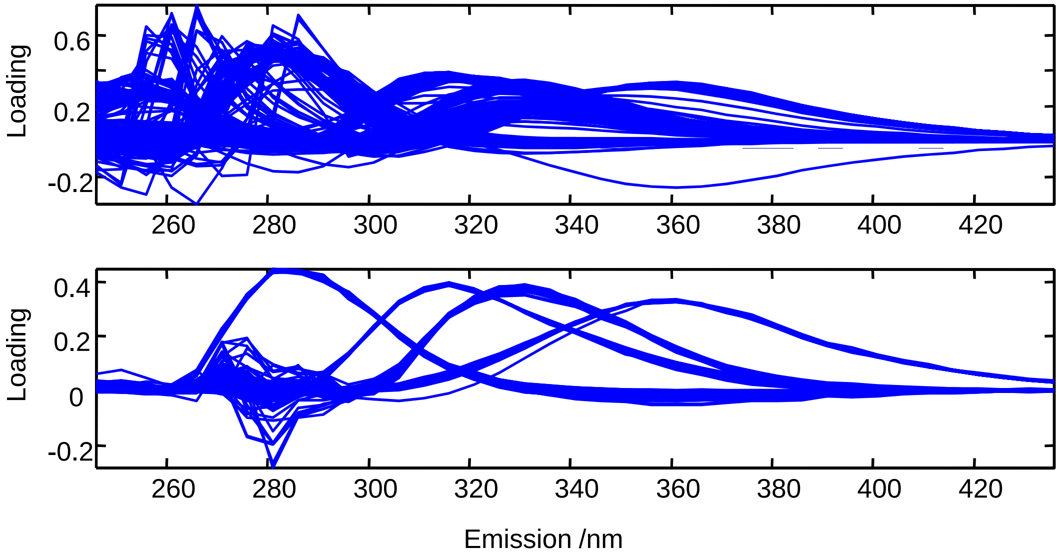
Least squares PARAFAC

MILES PARAFAC

Artifact

Bootstrapping a bit

Emission spectra from 100 resamplings



Least squares missing

"Maximum likelihood"

R. Bro, N. D. Sidiropoulos, and A. K. Smilde. Maximum likelihood fitting using simple least squares algorithms. J.Chemometrics, 2002

Handling shifts in data

- **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*

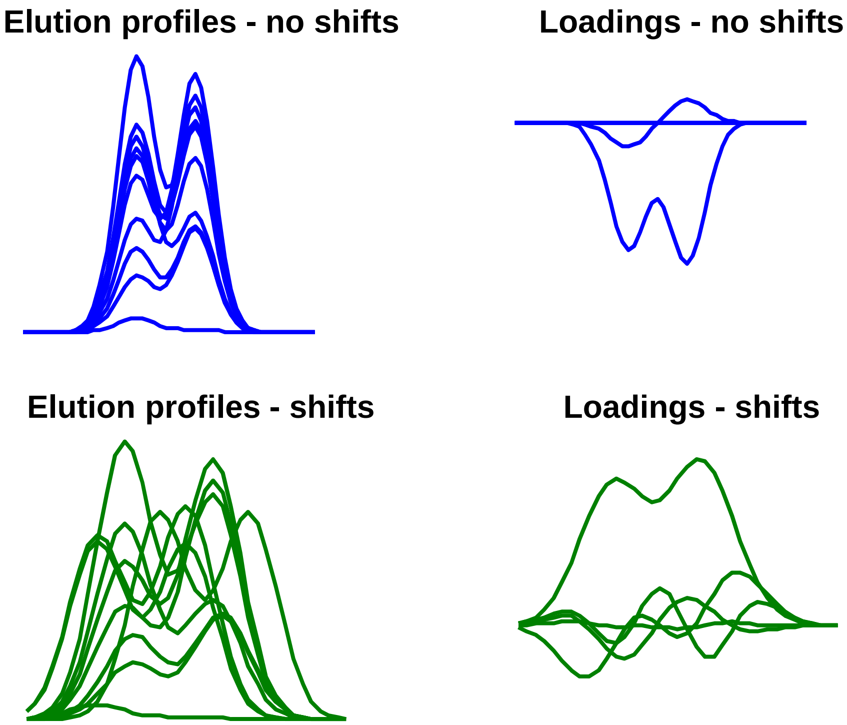
PARAFAC2 $\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}_{\boxed{k}}^T$ subject to $\mathbf{B}_k^T\mathbf{B}_k$ constant

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.

PARAFAC2 for shifted data

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



Elution profiles - no shifts

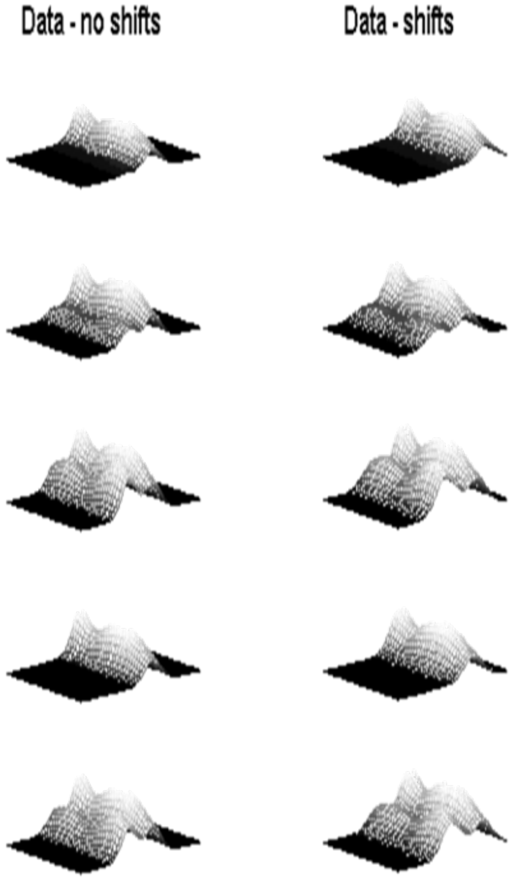
Loadings - no shifts

Elution profiles - shifts

Loadings - shifts

PARAFAC2 – new possibilities

- **Same data with spectral detection**
  - Difficult to see the shifts but exactly the same as before



PARAFAC2 – new possibilities

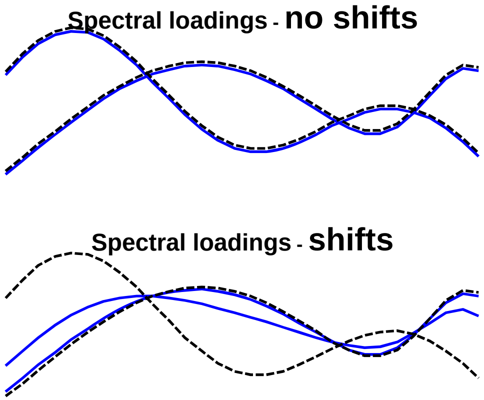
- **PARAFAC**
  - Works when no shifts but not with shifts

PARAFAC2

- Works regardless of shifts (in this case!)

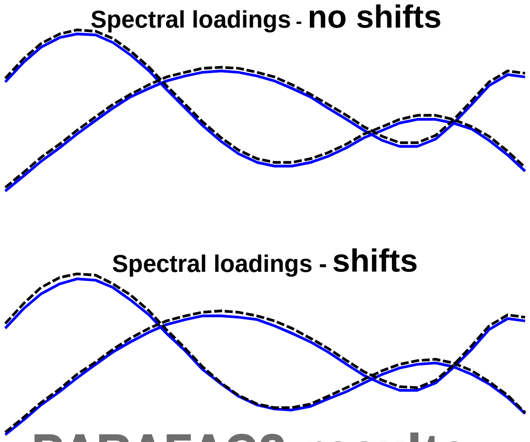
PARAFAC results

True spectra shown dashed



Spectral loadings - **no shifts**

Spectral loadings - **shifts**



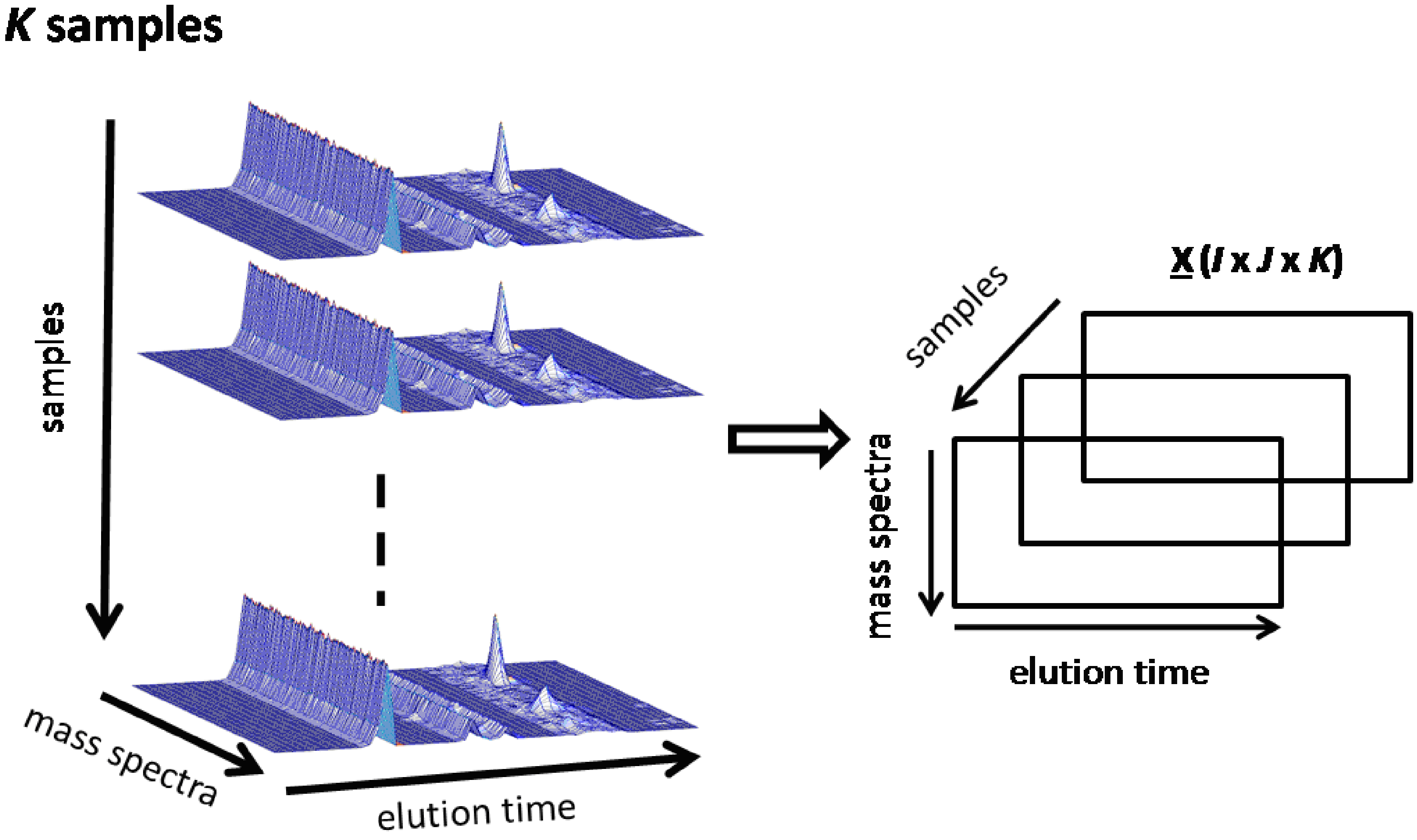
Spectral loadings - **no shifts**

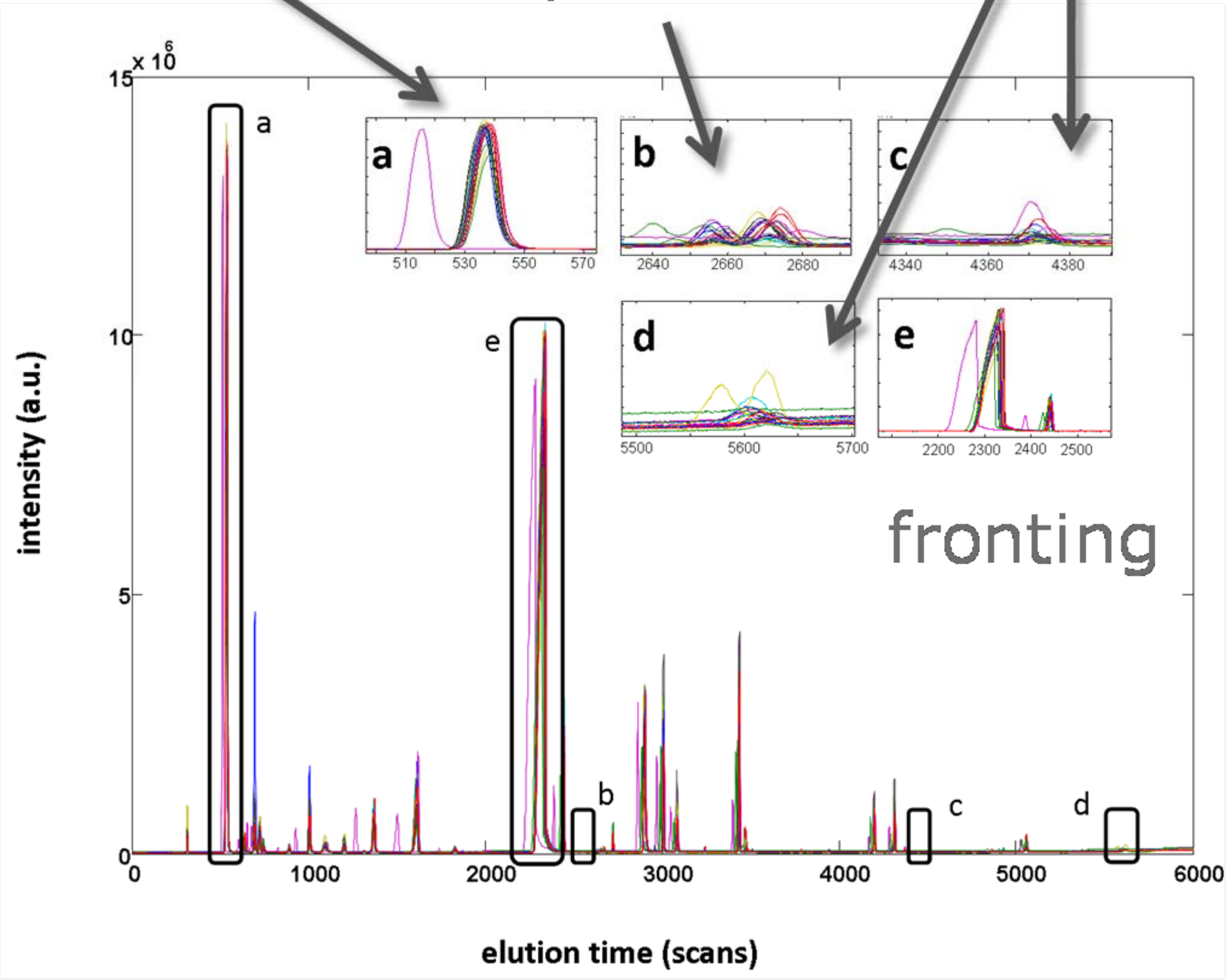
Spectral loadings - **shifts**

PARAFAC2 results

True spectra shown dashed

60 wine samples measured by GC-MS





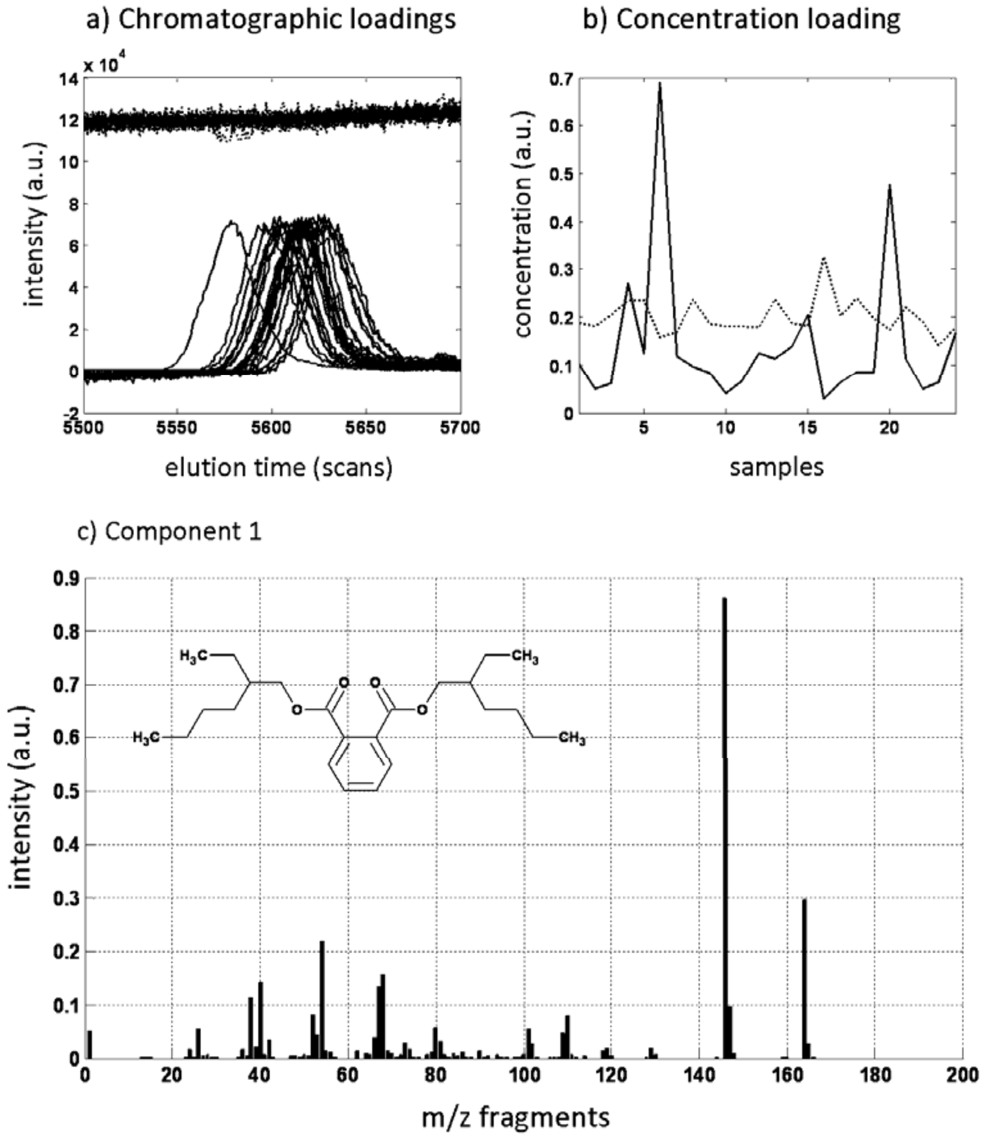
Weird shifts

Overlap and shifts

Low intensity

fronting

PARAFAC2 results



a) Chromatographic loadings

b) Concentration loading

c) Component 1

Tensor models provide

Mathematical chromatography
Huge noise reduction
Intuitive models (chemically)
Better handling of correlations
Robustness
...
But you need to know your data well – or be lucky

Still 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

csmr.ca.sandia.gov/~tgkolda/TensorToolbox/	Tensor Toolbox
www.eigenvector.com	Commercial software

Rasmus Bro
rb@life.ku.dk