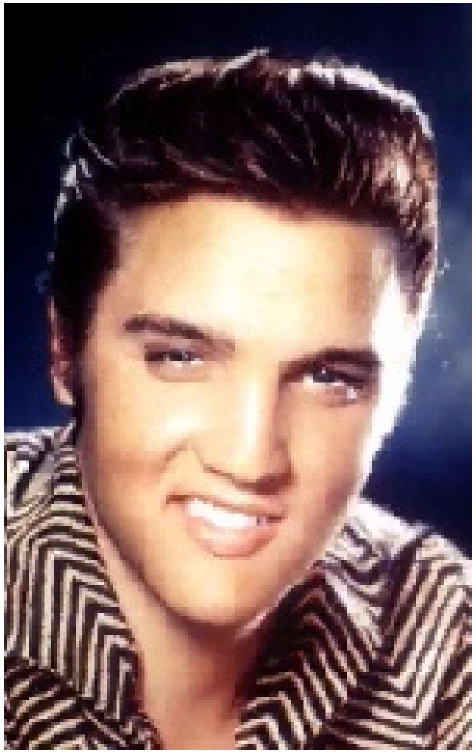
Multi-way analysis of bioinformatic data

Rasmus Bro
Chemometrics Group
Dept. Food Science
Royal Veterinary & Agricultural University (KVL)
rb@kvl.dk

Outline

- A little on how chemometrics uses visualization
- Why PARAFAC is so interesting
- Where it makes a difference

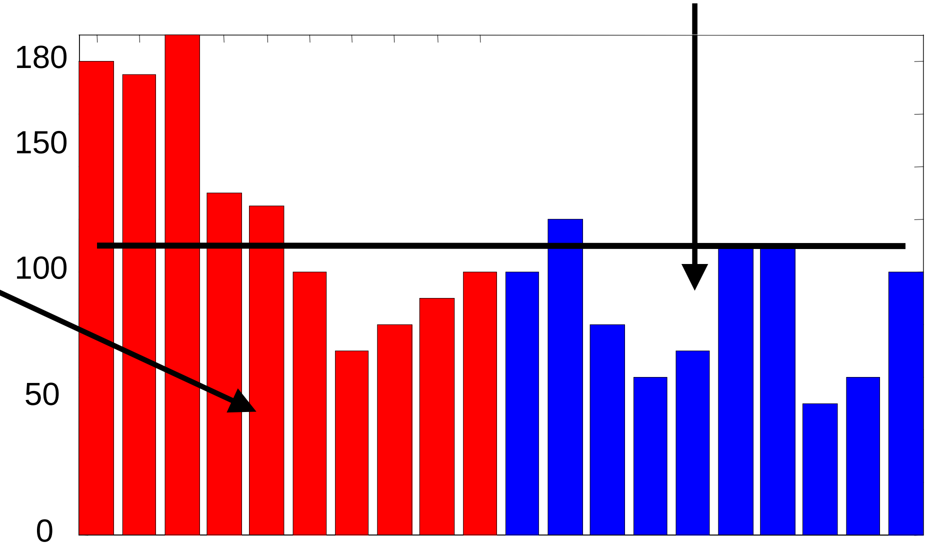
Human pattern recognition uses all available data



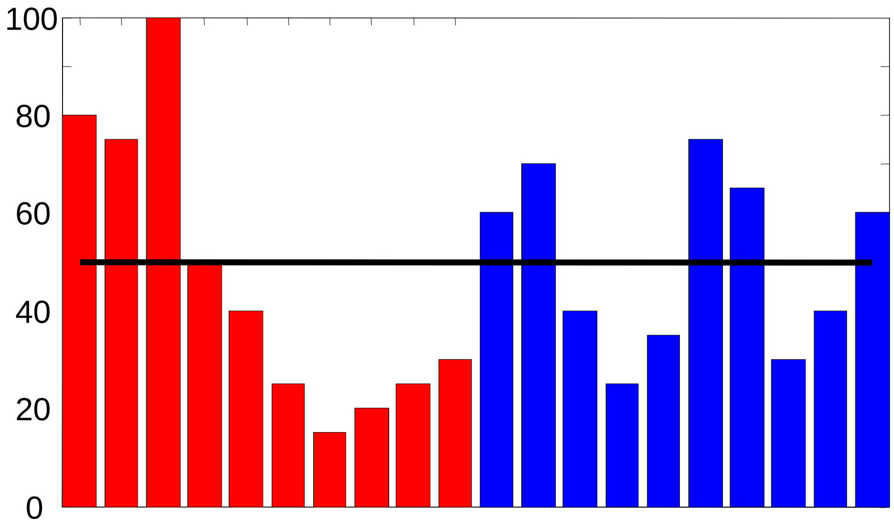
Single measurements / bar-plot

Height monkeys

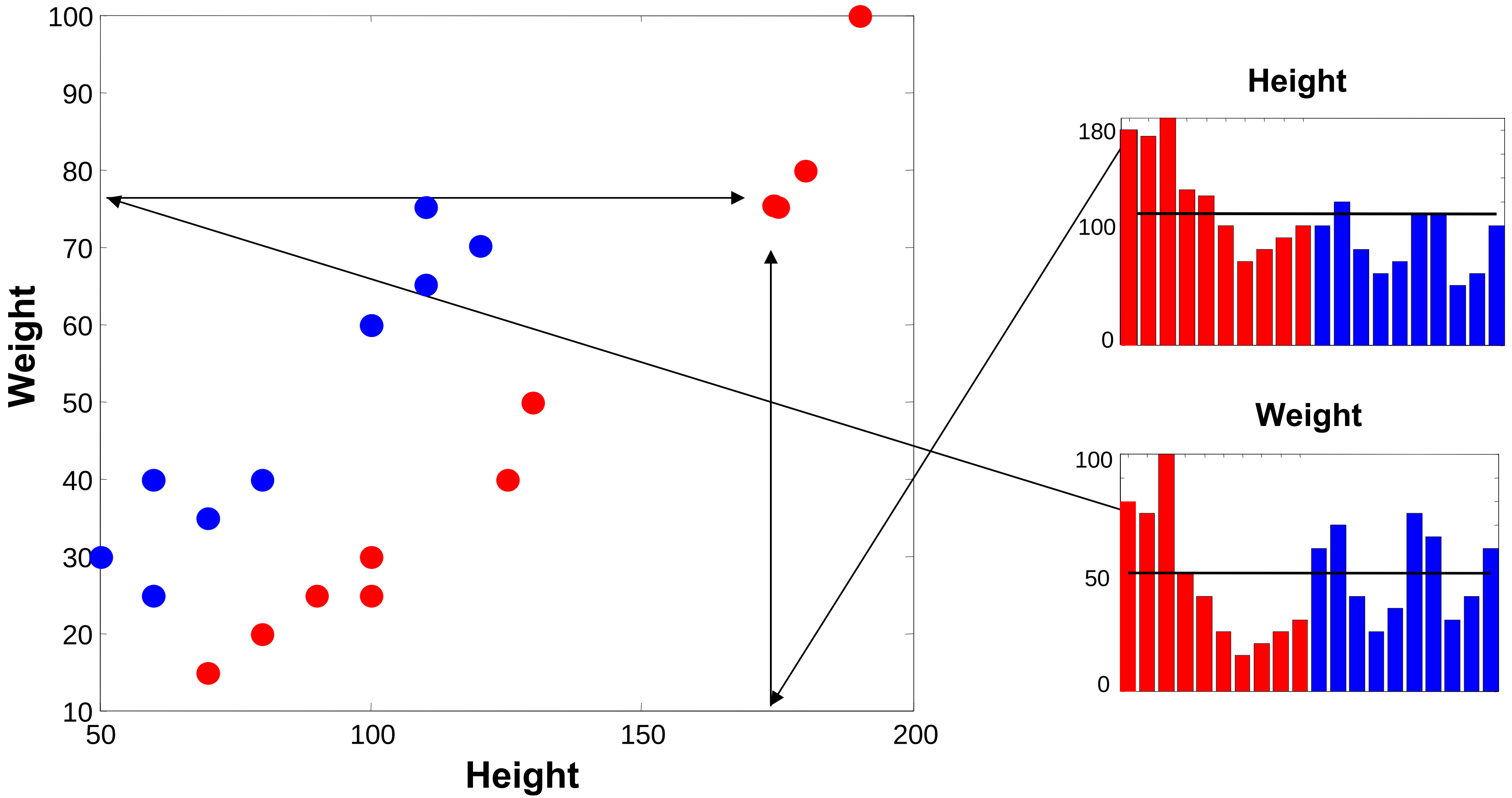
Height humans



Ditto weight



Pattern recognition



Principal component analysis movie

Removed. Find it at www.models.kvl.dk

Data analysis

Data ≠ Information ≠ Knowledge/Interpretation

7 variables

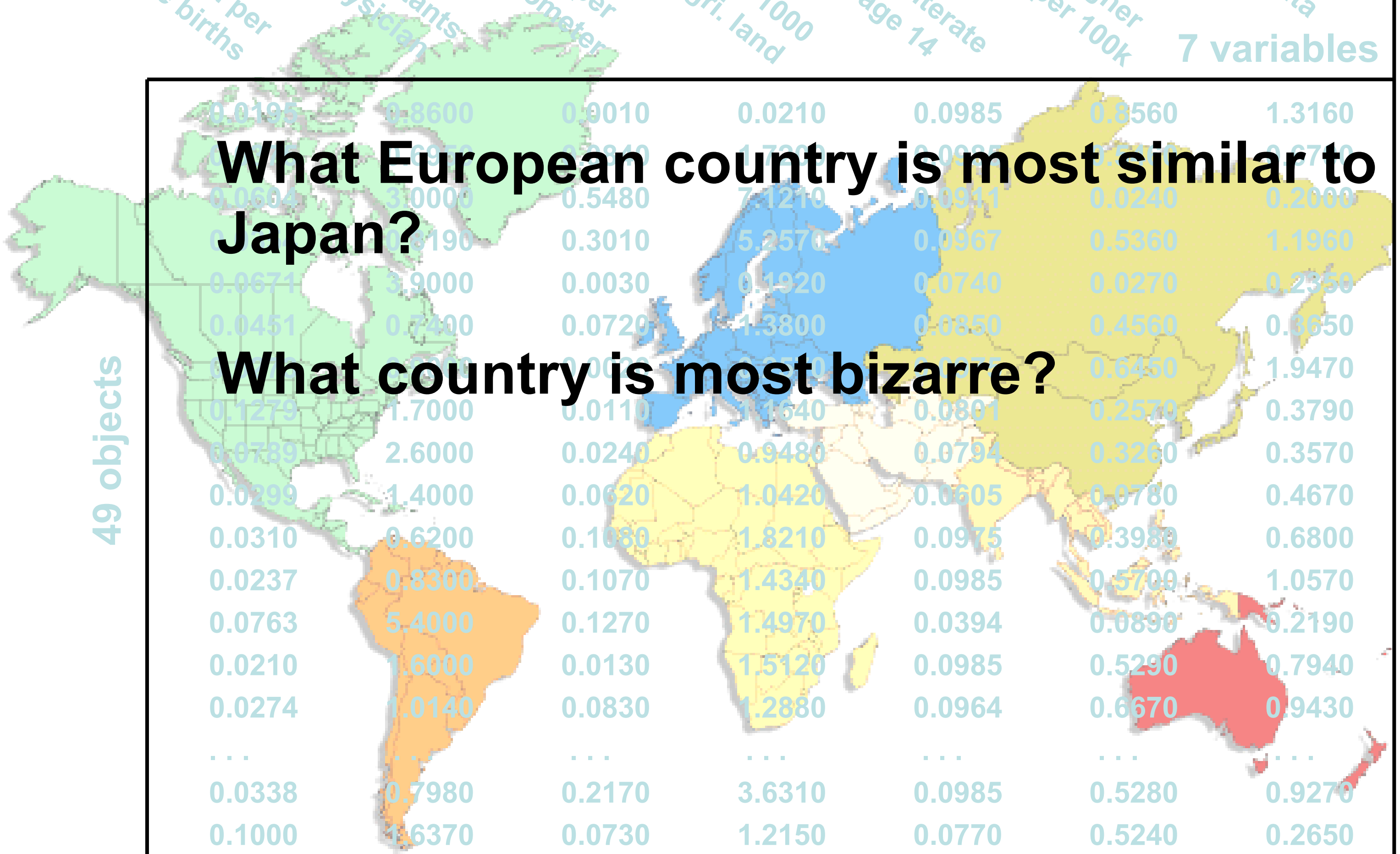
49 objects

countries	Infant death per 1000 live births	# of inhabitants per physician	population per square kilometer	population per 1000 hectares agri. land	percentage literate above age 14	# students higher education per 100k	GNP per capita
Austral	0.0195	0.8600	0.0010	0.0210	0.0985	0.8560	1.3160
Austria	[illegible]	0.6950	0.0840	1.7200	0.0985	0.5460	0.6700
Barbado	[illegible]	3.0000	0.5480	7.1210	0.0911	0.0240	0.2000
Belgium	[illegible]	[illegible]	0.3010	5.2570	0.0967	0.5360	1.1960
Brit Gu	0.0671	[illegible]	0.0030	0.1920	0.0740	0.0270	0.2350
Bulgari	0.0451	[illegible]	[illegible]	1.3800	0.0850	0.4560	0.3650
Canada	0.0273	0.9000	[illegible]	0.2570	0.0975	0.6450	1.9470
Chile	0.1279	1.7000	[illegible]	[illegible]	0.0801	0.2570	0.3790
Costa R	0.0789	2.6000	[illegible]	[illegible]	0.0794	0.3260	0.3570
Cyprus	0.0299	1.4000	0.0620	[illegible]	[illegible]	0.0780	0.4670
Czechos	0.0310	0.6200	0.1080	[illegible]	[illegible]	0.3980	0.6800
Denmark	0.0237	0.8300	0.1070	1.4340	[illegible]	[illegible]	1.0570
El Salv	0.0763	5.4000	0.1270	1.4970	[illegible]	[illegible]	0.2190
Finland	0.0210	1.6000	0.0130	1.5120	0.0985	[illegible]	[illegible]
France	0.0274	1.0140	0.0830	1.2880	0.0964	[illegible]	[illegible]
. . .	. . .	. . .	. . .	. . .	. . .	. . .	
West Ge	0.0338	0.7980	0.2170	3.6310	0.0985	0.5280	[illegible]
Yugosla	0.1000	1.6370	0.0730	1.2150	0.0770	0.5240	0.2650

x1000

Gunst & Mason (1980) Regression analysis and its applications: A data-oriented approach, NY, Marcel Dekker, p. 358

Incredibly simple questions



7 variables

49 objects

countries	Infant death per 1000 live births	# of inhabitants per physician	population per square kilometer	population per 1000 hectares agri. land	percentage literate above age 14	# students higher education per 100k	GNP per capita
Austral	0.0195	0.8600	0.0010	0.0210	0.0985	0.8560	1.3160
Austria	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Barbado	0.0604	3.0000	0.5480	7.1210	0.0911	0.0240	0.2000
Belgium	[illegible]	0.8190	0.3010	5.2570	0.0967	0.5360	1.1960
Brit Gu	0.0671	3.9000	0.0030	0.1920	0.0740	0.0270	0.2350
Bulgari	0.0451	0.7400	0.0720	1.3800	0.0850	0.4560	0.3650
Canada	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.6450	1.9470
Chile	0.1279	1.7000	0.0110	1.1640	0.0801	0.2570	0.3790
Costa R	0.0789	2.6000	0.0240	0.9480	0.0794	0.3260	0.3570
Cyprus	0.0299	1.4000	0.0620	1.0420	0.0605	0.0780	0.4670
Czechos	0.0310	0.6200	0.1080	1.8210	0.0975	0.3980	0.6800
Denmark	0.0237	0.8300	0.1070	1.4340	0.0985	0.5700	1.0570
El Salv	0.0763	5.4000	0.1270	1.4970	0.0394	0.0890	0.2190
Finland	0.0210	1.6000	0.0130	1.5120	0.0985	0.5290	0.7940
France	0.0274	1.0140	0.0830	1.2880	0.0964	0.6670	0.9430
. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
West Ge	0.0338	0.7980	0.2170	3.6310	0.0985	0.5280	0.9270
Yugosla	0.1000	1.6370	0.0730	1.2150	0.0770	0.5240	0.2650

x1000

What European country is most similar to Japan?

What country is most bizarre?

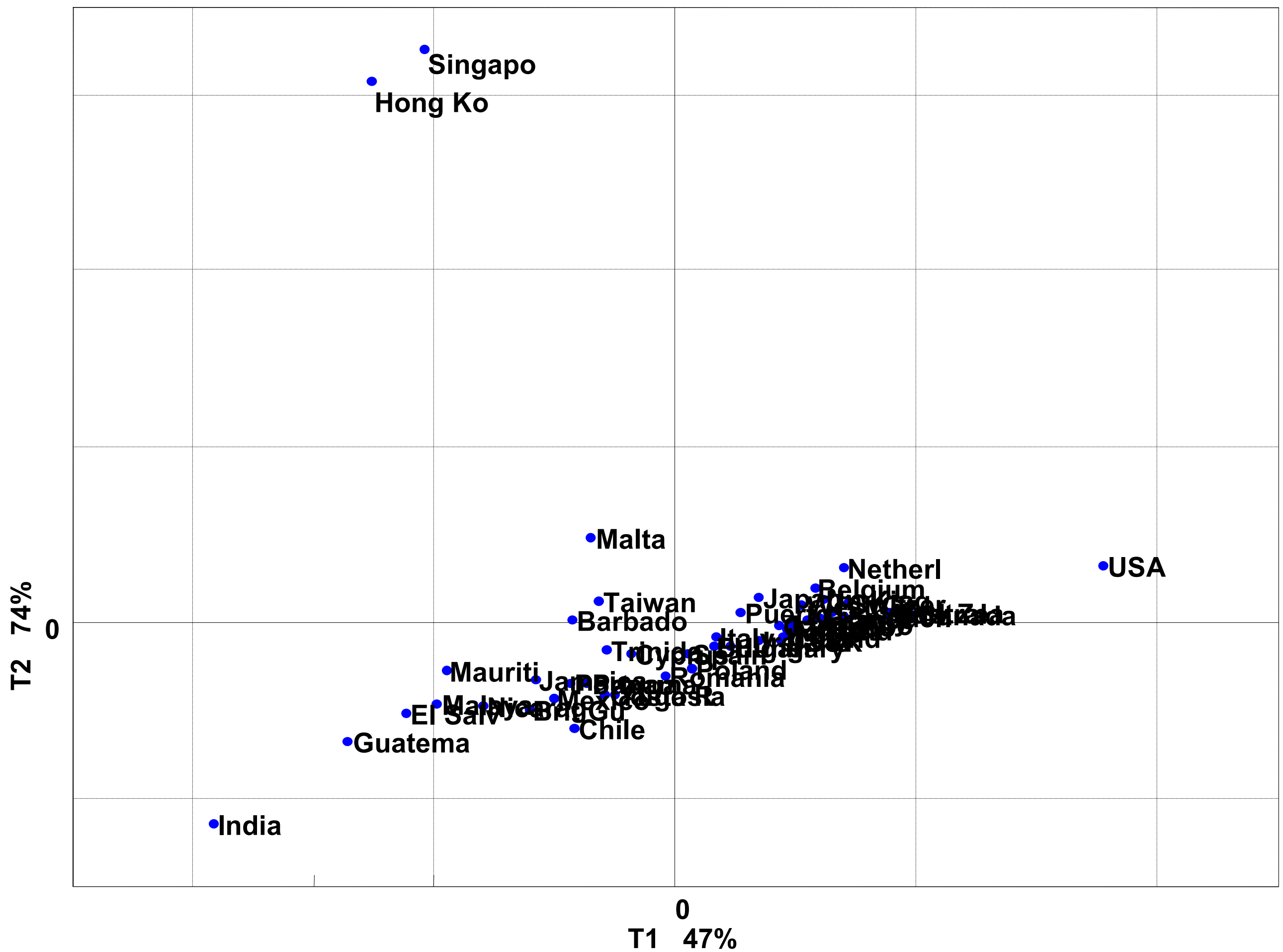
Principal Component Analysis

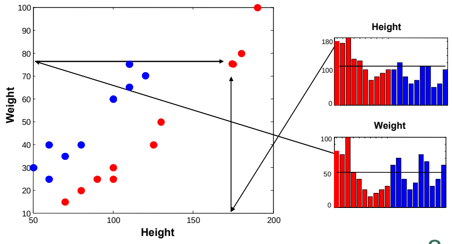
Outliers are easily spotted in score scatter plot

$$X = TP' + E$$

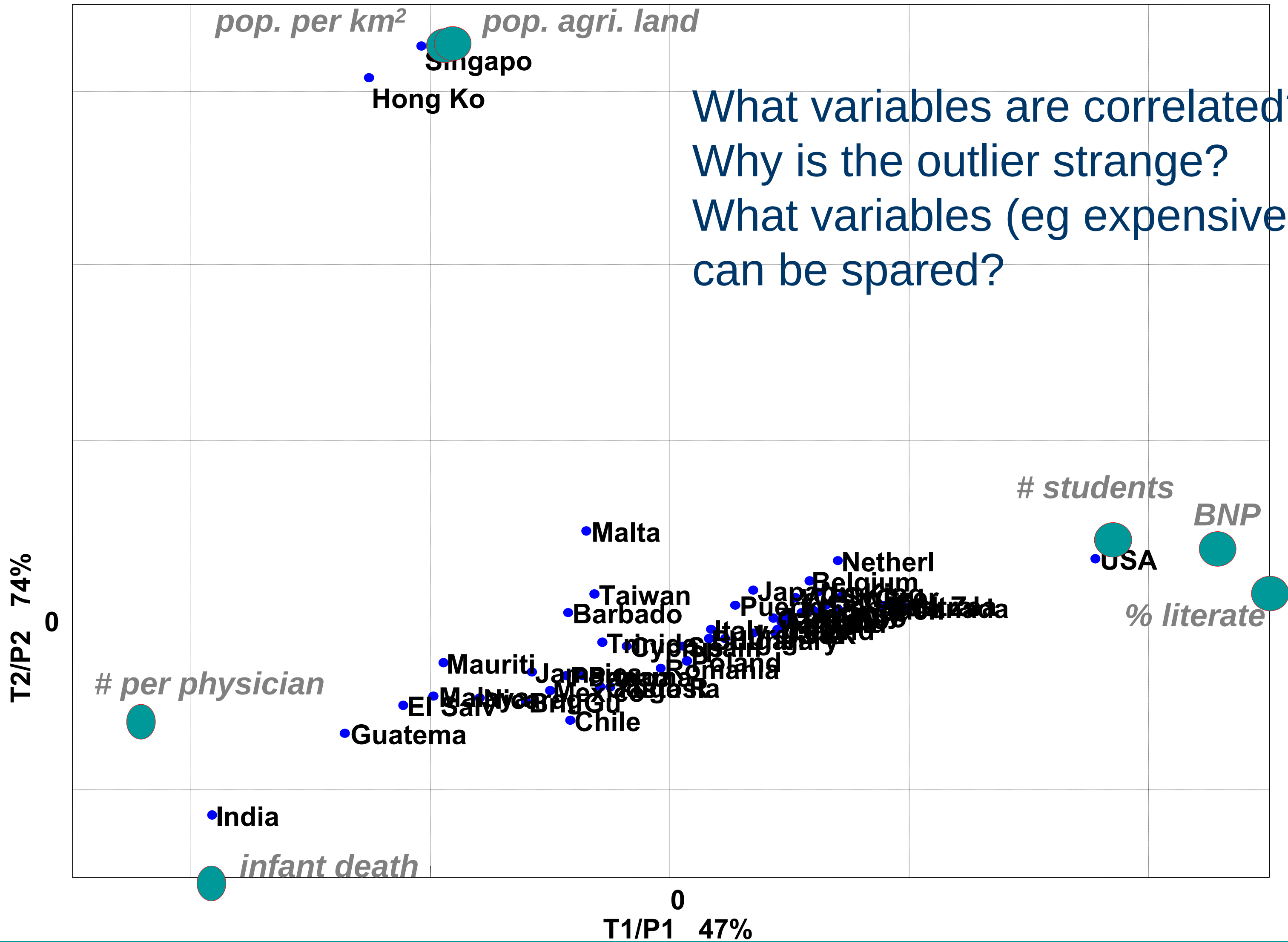
Scores

Loadings



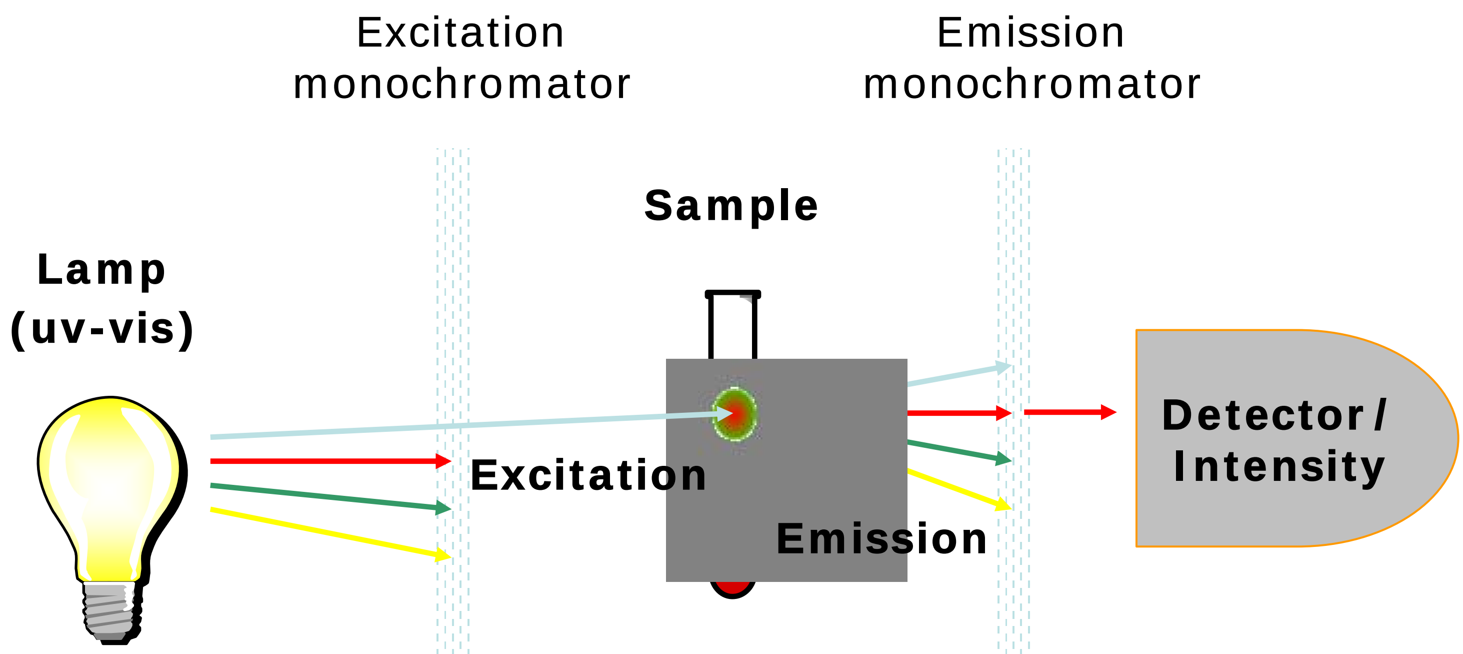


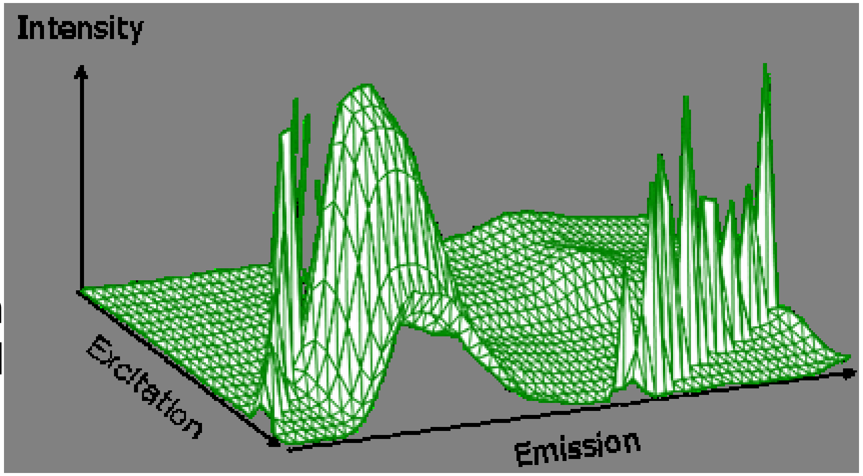
Principal Component Analysis
Why – Add loadings – bi-plot?



What variables are correlated?
Why is the outlier strange?
What variables (eg expensive) can be spared?

Fluorescence spectroscopy





Excitation-emission matrix – a chemical fingerprint

PARAFAC movie

Removed. Find it at www.models.kvl.dk

PARAFAC movie

Removed. Find it at www.models.kvl.dk

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

$$\mathbf{X}_k = \mathbf{A}\mathbf{D}_k\mathbf{B}' + \mathbf{E}_k$$

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

PARAFAC - algorithm

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

Why ALS?

Simple

Extends to N-way

Handles missing

Handles ML fitting

Constraints:

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

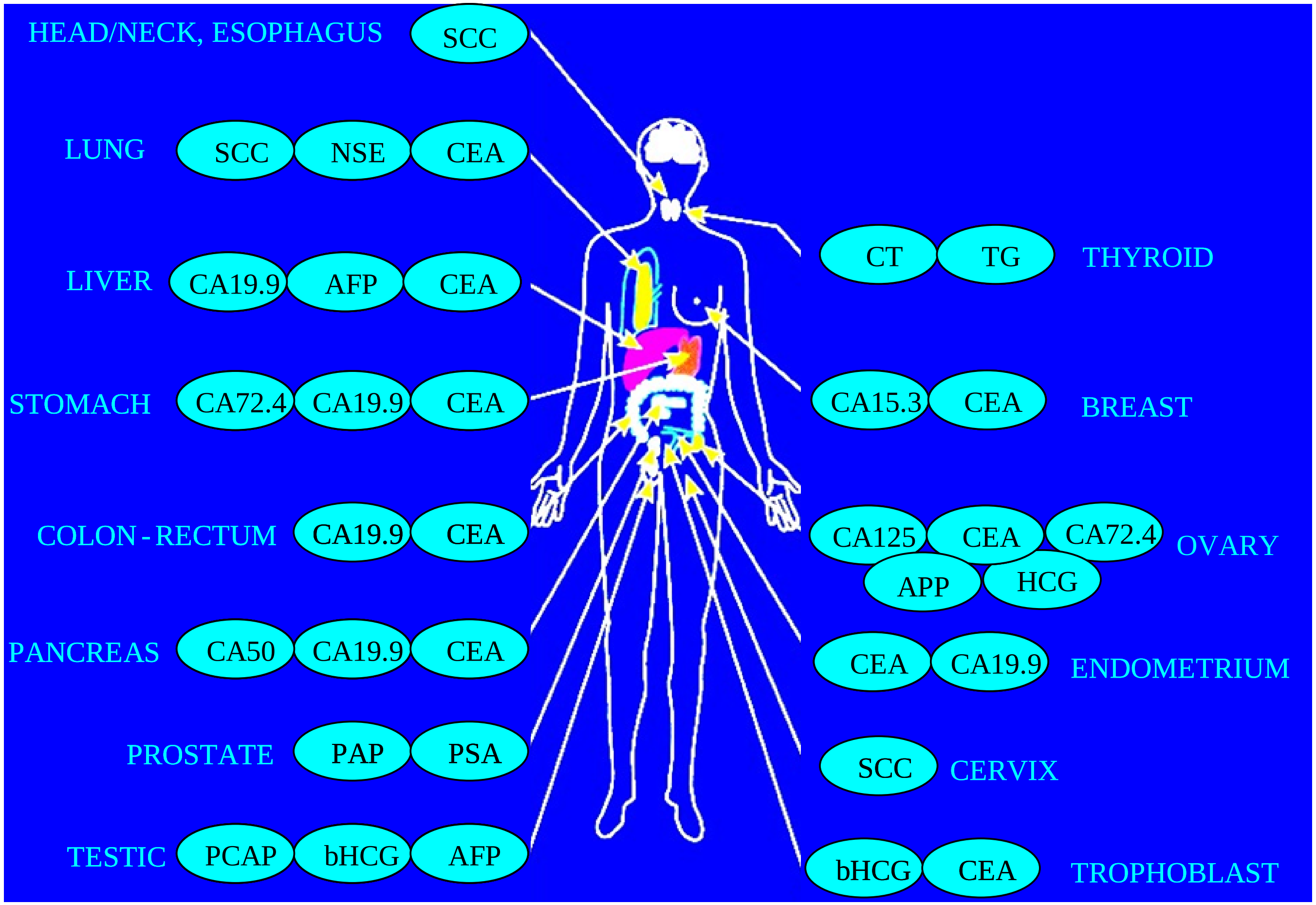
NMF dates back

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

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

Cancer diagnostics

Traditional Approach Biomarkers

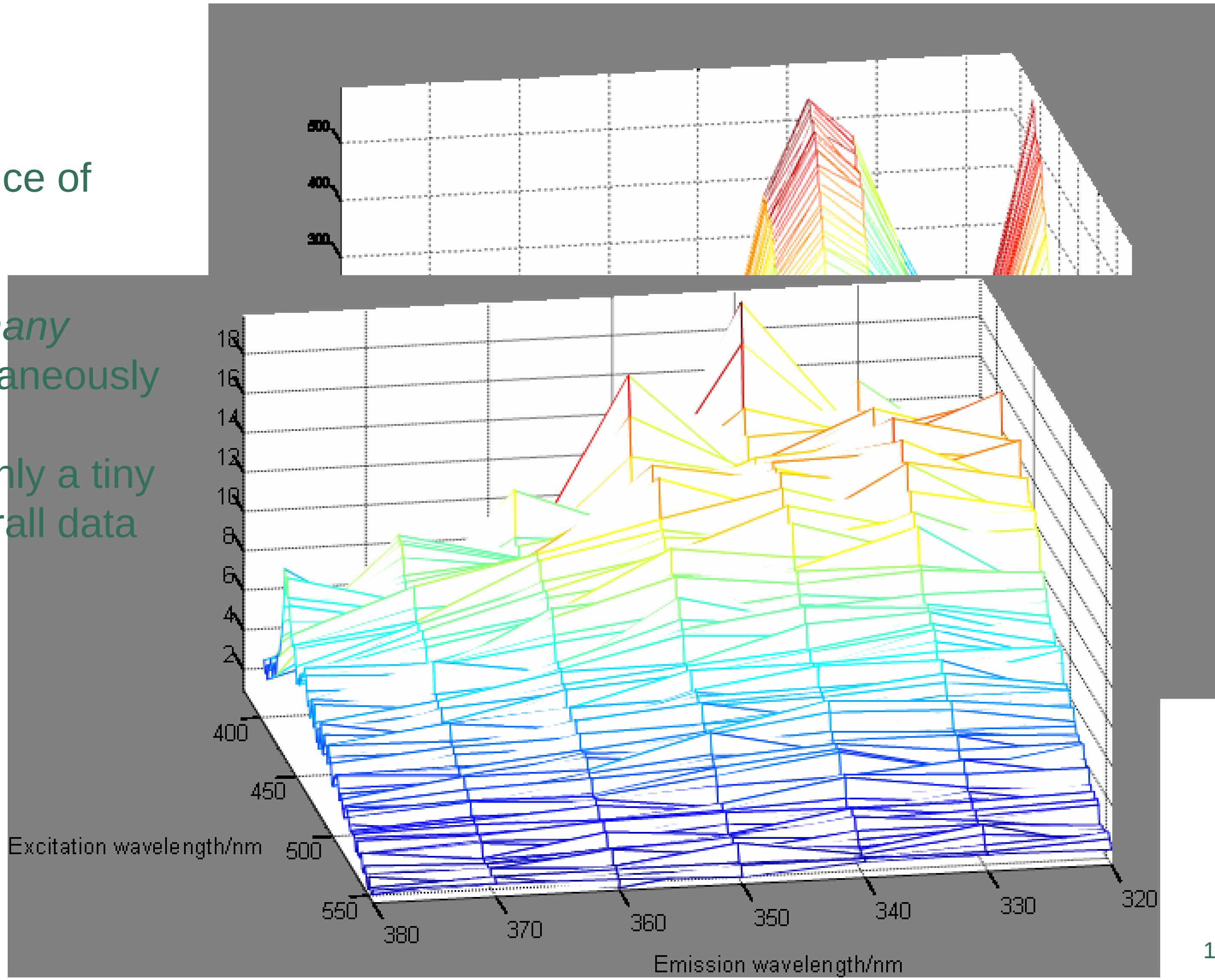


Cancer diagnostics

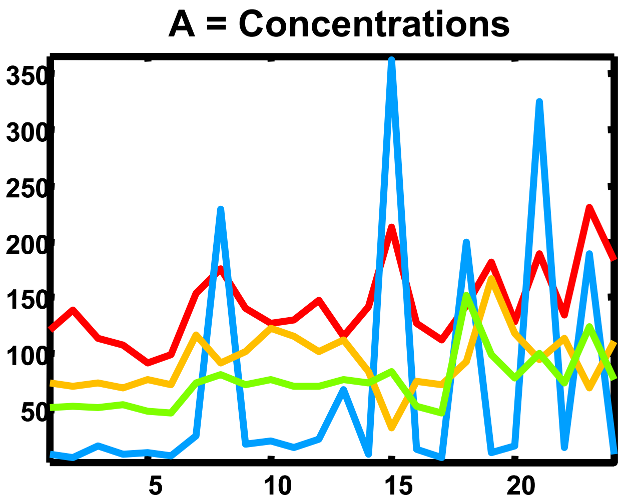
Alternative
Use fluorescence of blood samples

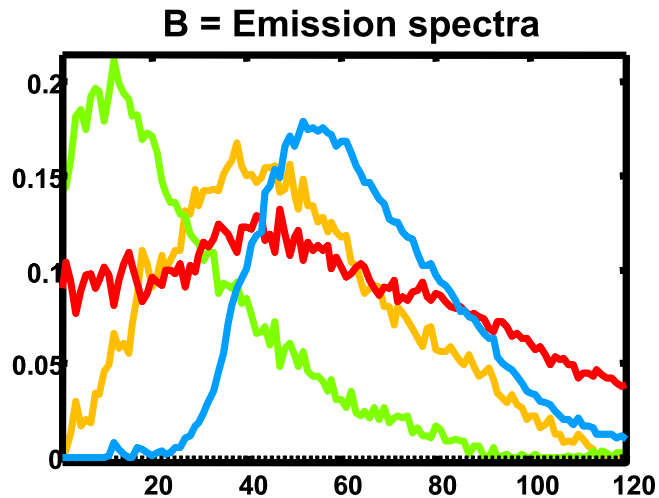
I.e. measure *many* markers simultaneously

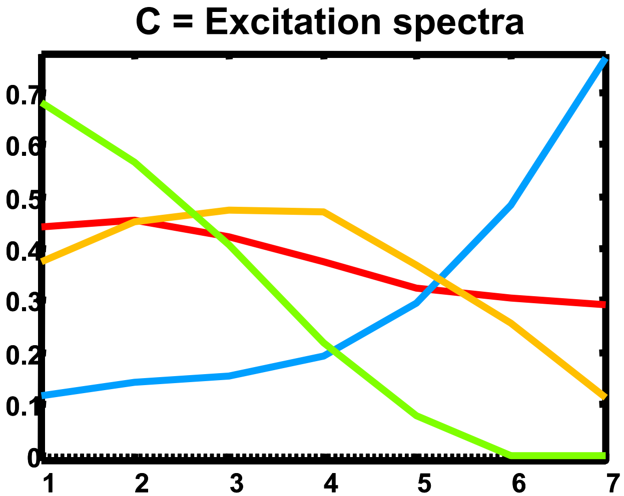
Here we use only a tiny part of the overall data



Cancer diagnostics

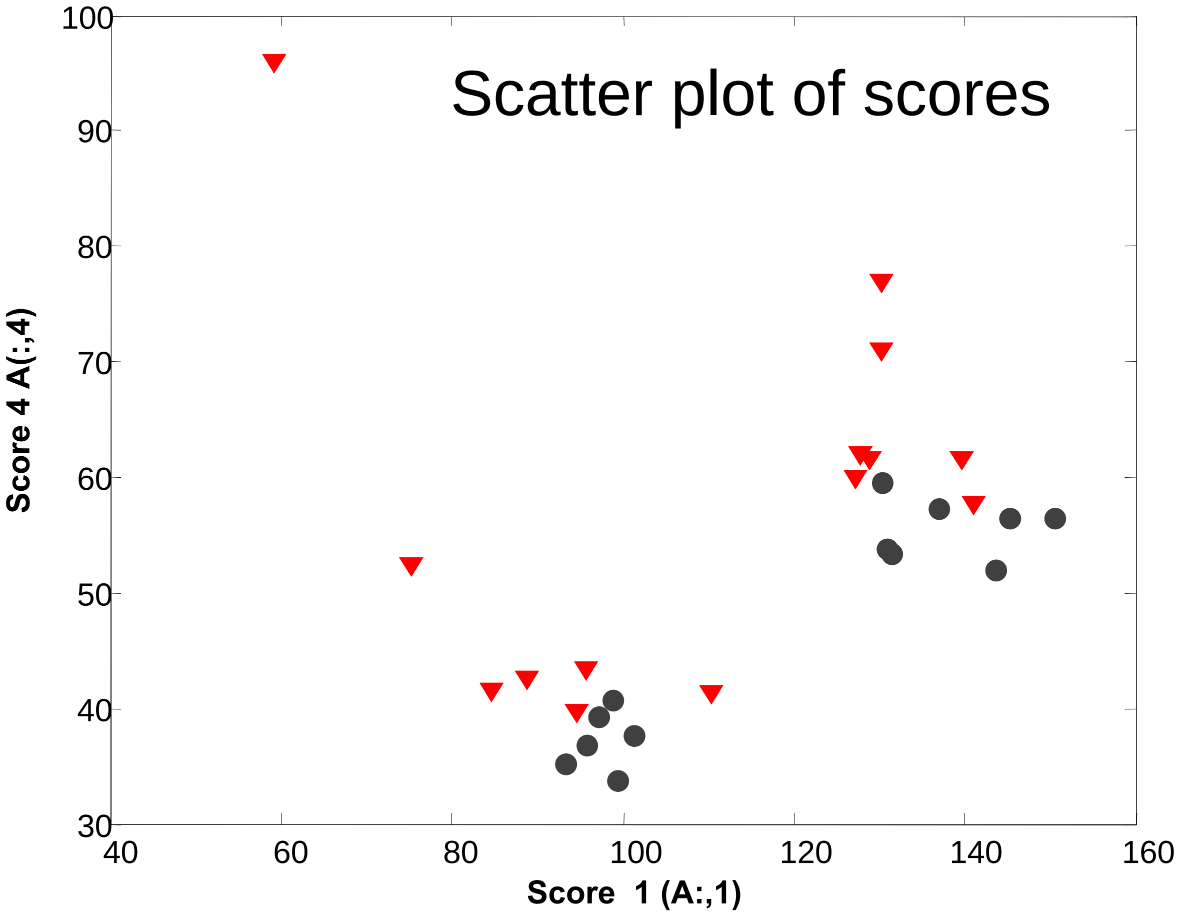






$$x_{ijk} = \sum_{f=1}^{F} a_{if} b_{jf} c_{kf} + e_{ijk}$$

Cancer diagnostics

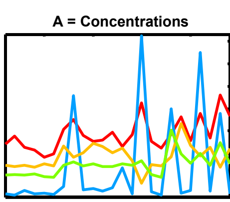


Tumor

Control

Grouping according to:
- Disease
- Something else.

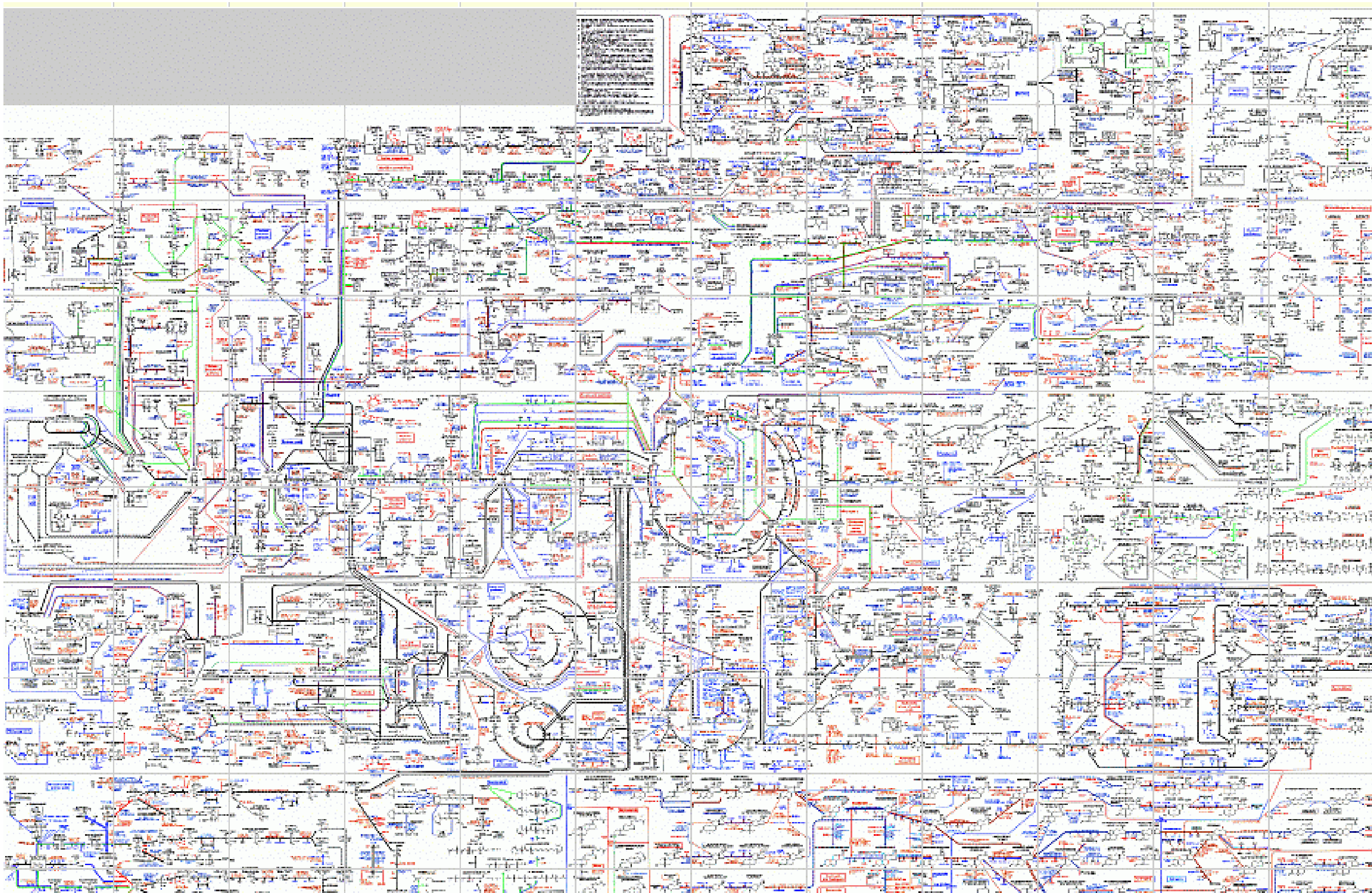
Chemically interpretable on a fluorescence level leading to understanding, biomarkers, etc.



Multi-way analysis in metabonomics

Metabonomics “the systematic study of the unique chemical fingerprints that specific cellular processes leave behind”.

A simplified view on metabolic pathways



From Roche

Toxic study

Typical data

Control (5 rats)
Low dose (5 rats)
High dose (5 rats)

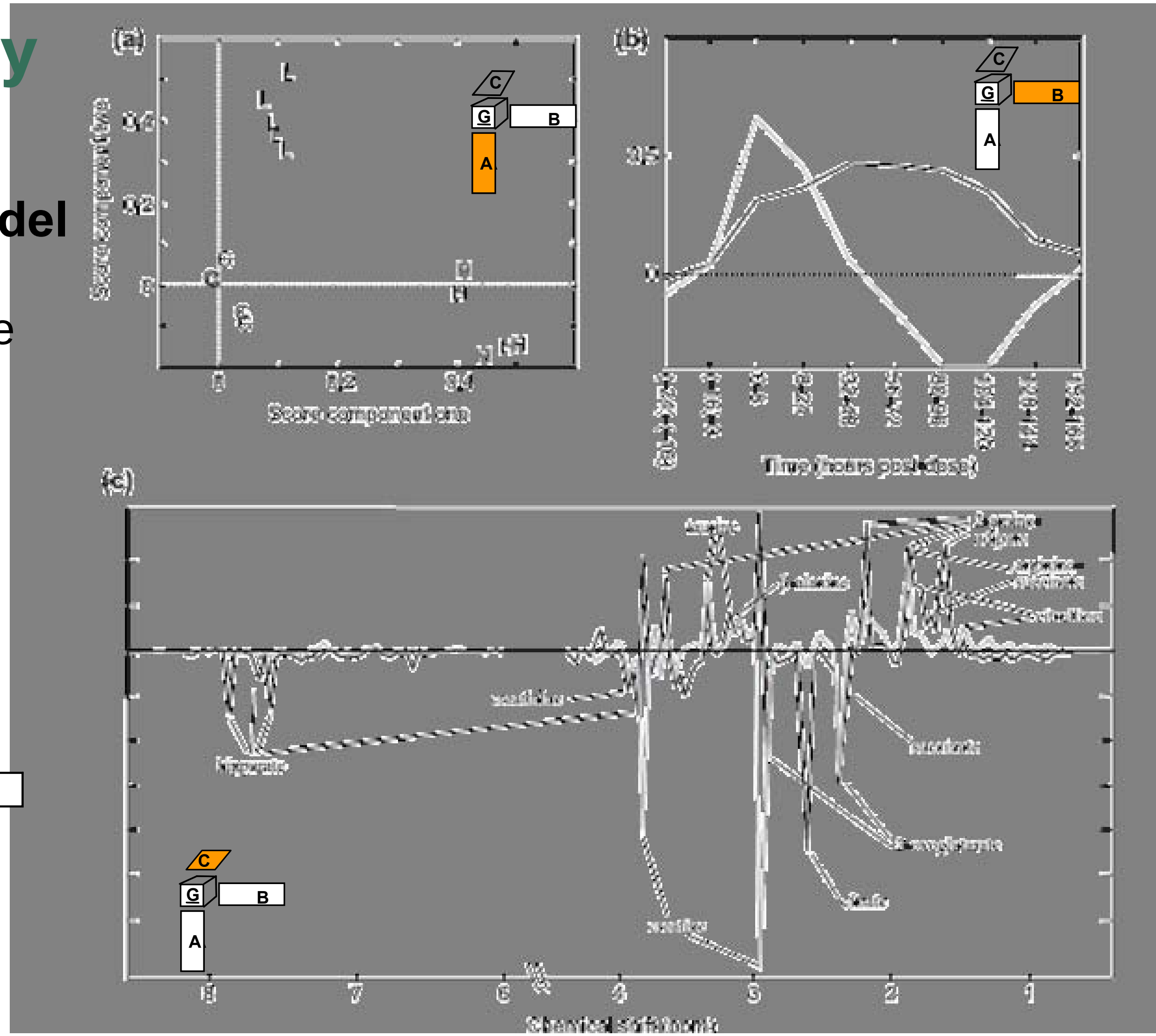

ELSEVIER

Chemometrics and Intelligent Laboratory Systems

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

Marianne Dyrby, Dorrit Baunsgaard, Rasmus Bro, Søren Balling Engelsen

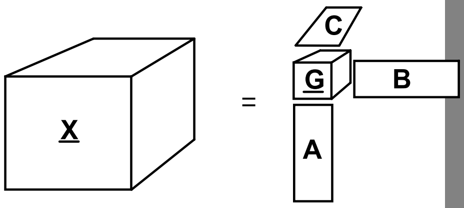
Toxic study



Exploratory Tucker model

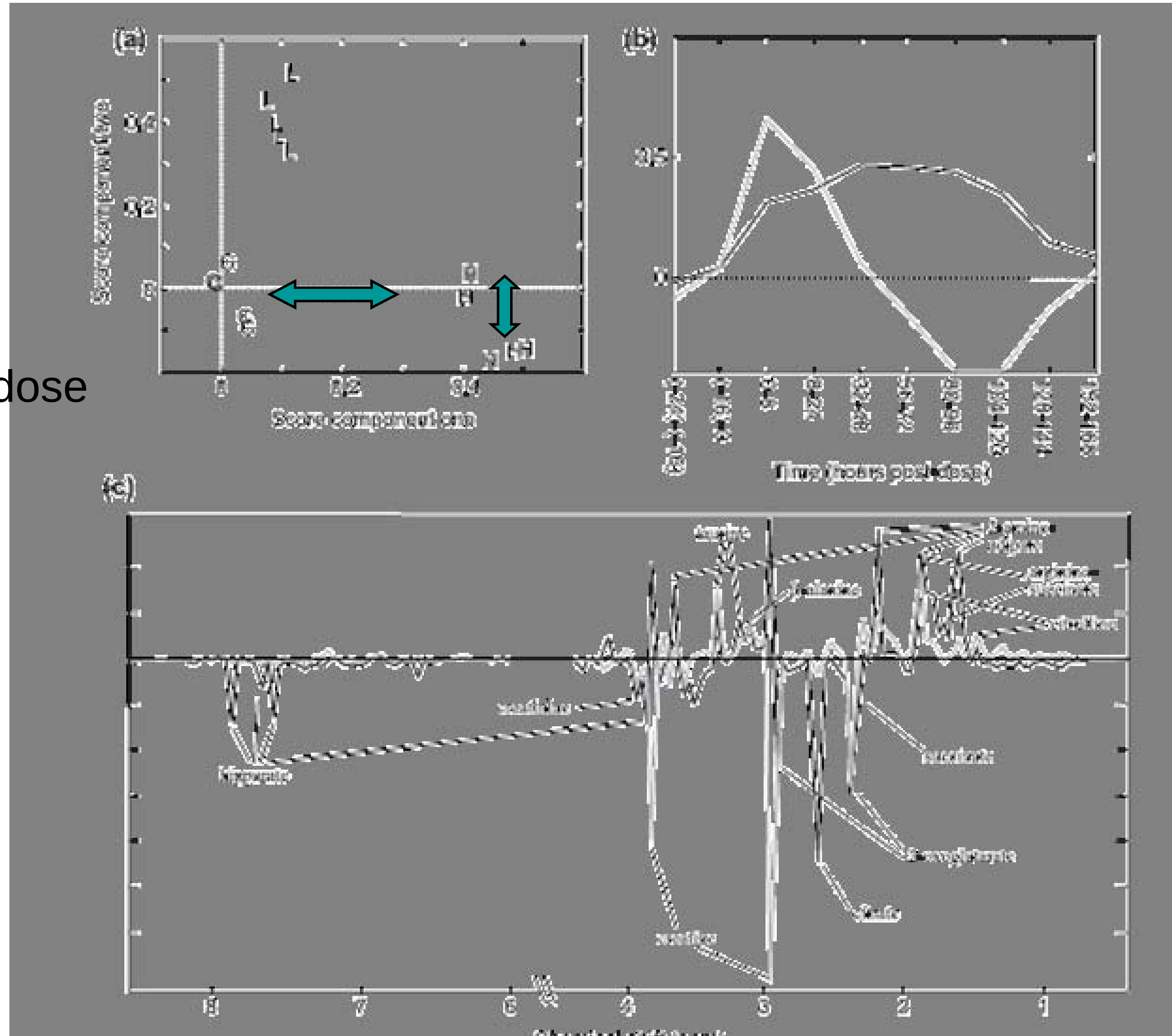
Unlike unfolding/flattening:

Shows recovery of low-dose rats as well as an early response to taurine and creatine.



Toxic study

Hydrazine study



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

Toxic study

Multilevel component analysis of time-resolved metabolic fingerprinting data

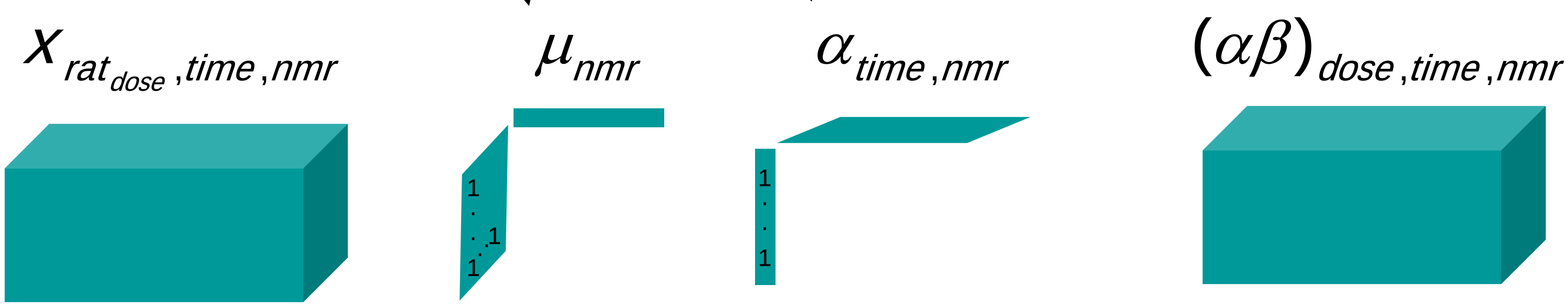
Jeroen J. Jansen[a], Huub C.J. Hoefsloot[a], Jan van der Greef[b,c], Marieke E. Timmerman[d], Age K. Smilde[a,b,*]

Further analysis

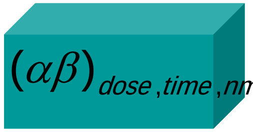
Use ANOVA-SCA (simultaneous component analysis)

ANOVA

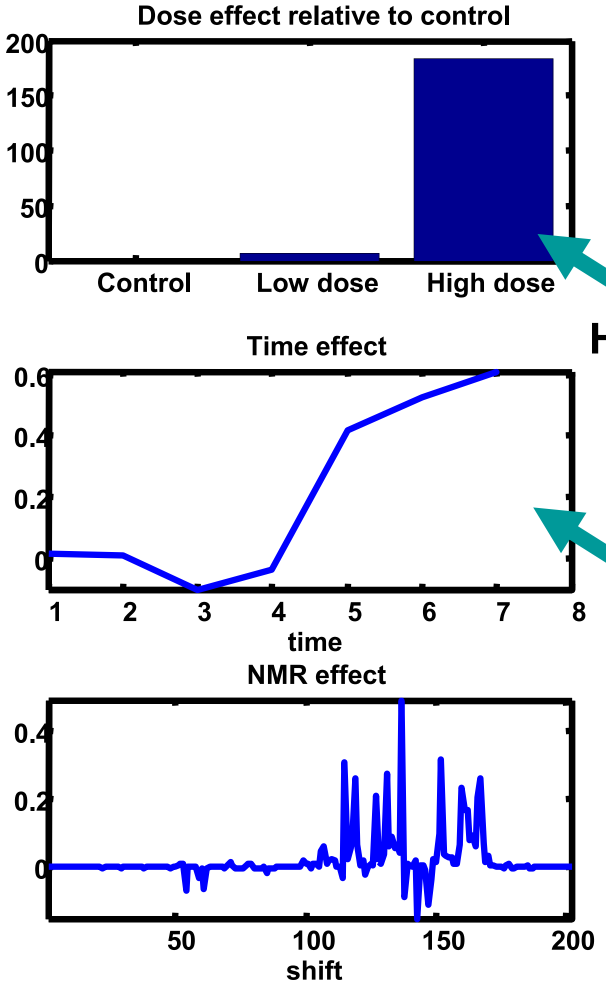
$$x_{rat_{dose},time,nmr} = \mu_{nmr} + \alpha_{time,nmr} + (\alpha\beta)_{dose,time,nmr} + e_{rat_{dose},time,nmr}$$



PARAFAC on dose / time-effect



Separate the effect into a shock

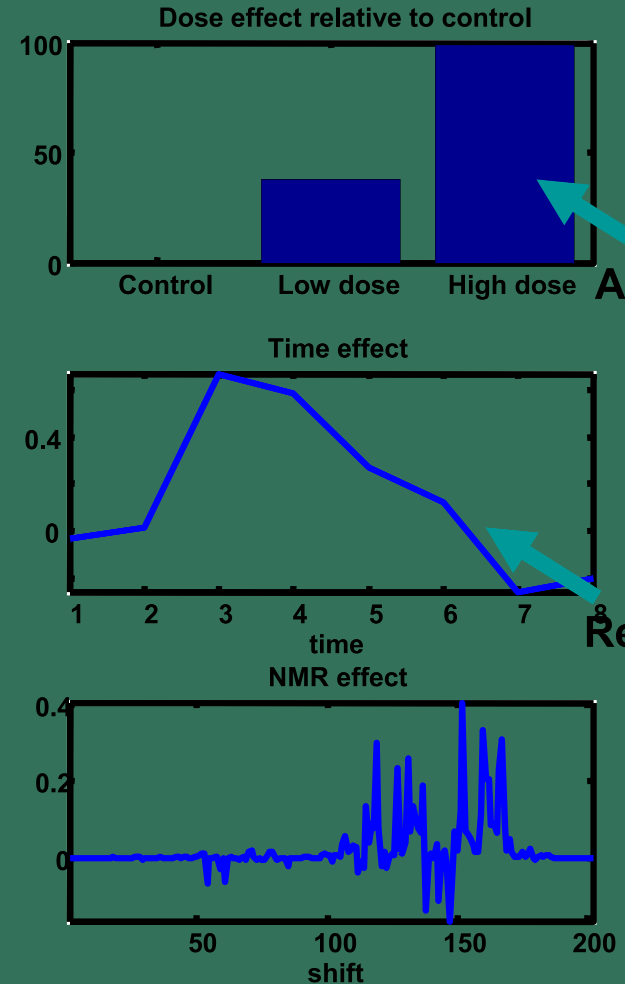


High-dose on

Irreversible

PARAFAC factor 1

and a reversible effect



All-dose linear

Reversible

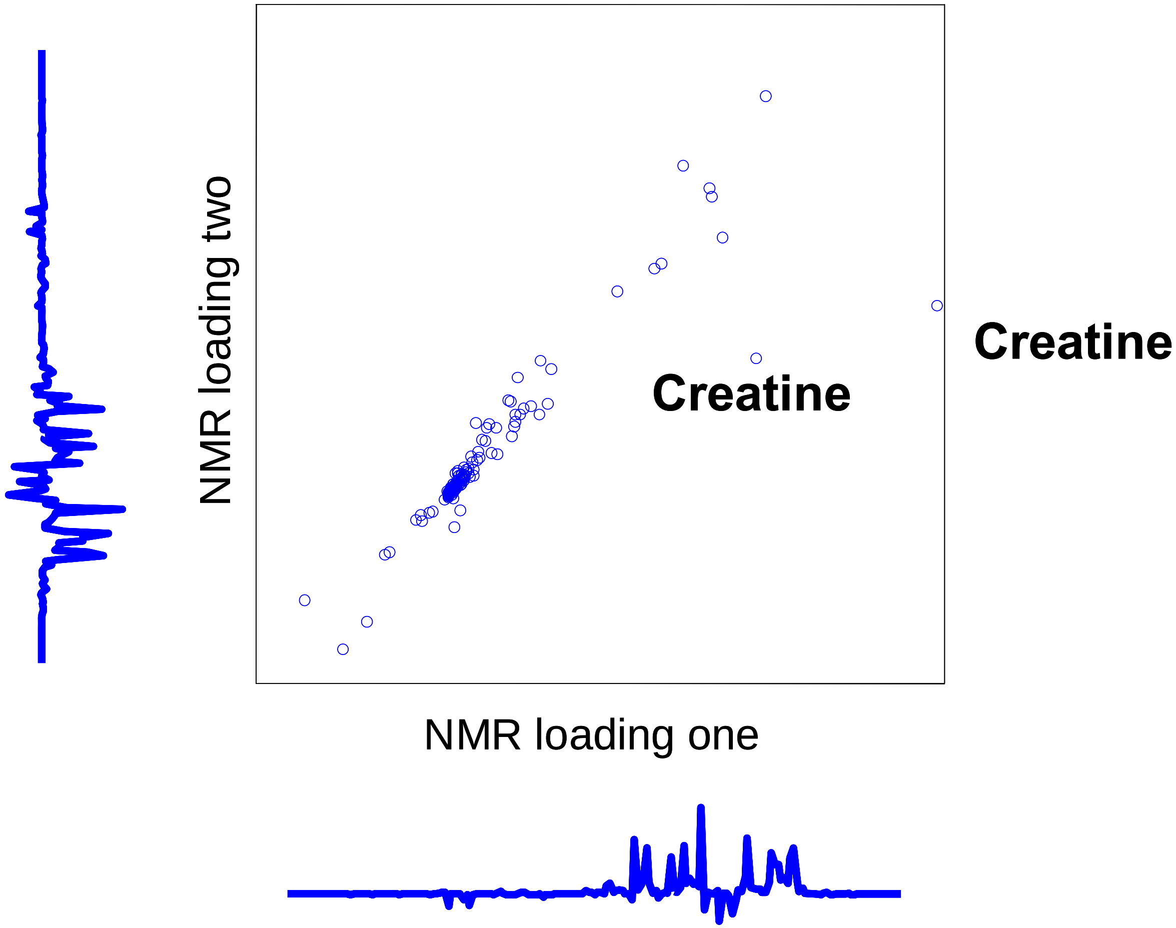
PARAFAC factor 2

Toxic study

$(\alpha\beta)_{dose,time,nmr}$

Difference between reversible and irreversible effect

Creatine indicating chronic kidney damage



Concluding remarks

Many interesting solutions using tensor approaches

Uniqueness

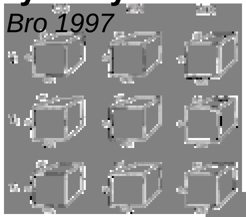
- Pure spectra
- Pure profiles
- Pure concentrations
- Pure magic!!!

=

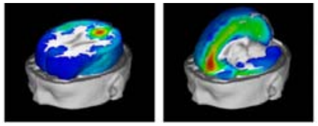
- Mathematical chromatography

Other Examples

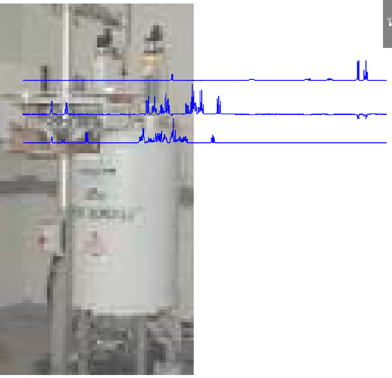
5-way analysis
Bro 1997



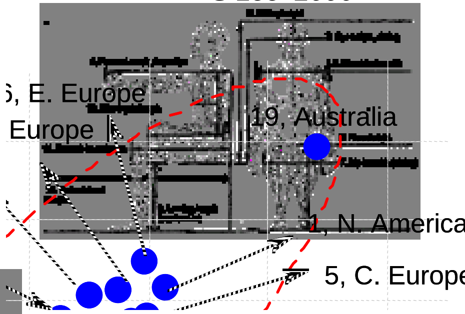
EEG
Miwakeichi et al 2004



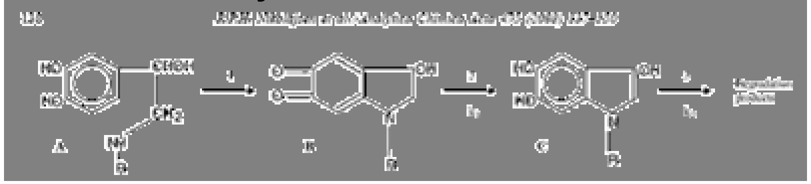
DOSY
Toft et al 2004



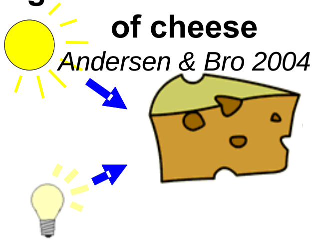
Anthropometry
S Lee 2006



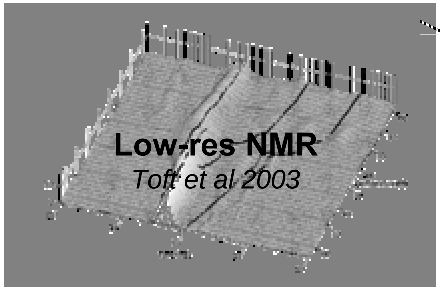
Quantify catecholamine in urine



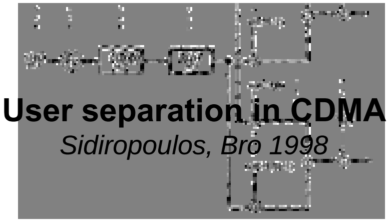
Light-induced oxidation of cheese
Andersen & Bro 2004



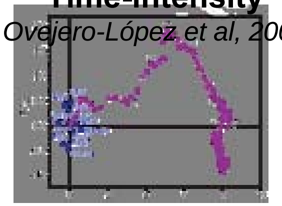
Low-res NMR
Toft et al 2003



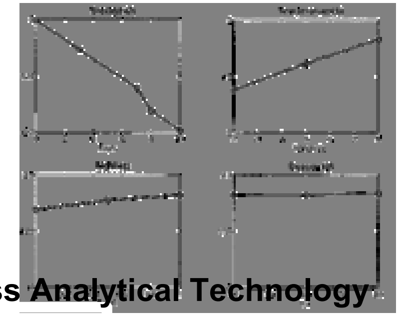
User separation in CDMA
Sidiropoulos, Bro 1998



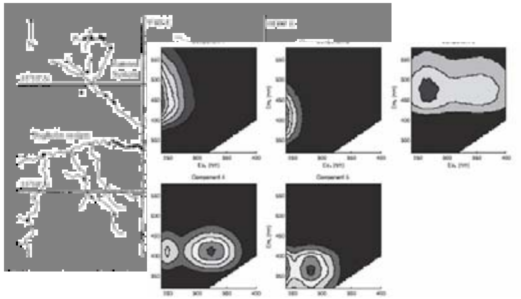
Time-intensity
Ovejero-López et al, 2004



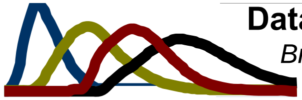
Generalized ANOVA
Bro & Jakobsen 1996, 2002



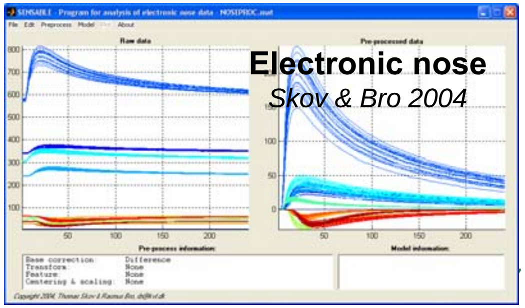
Tracing DOM
Stedmon, Markager, Bro 2003



Process Analytical Technology
Datamining
Bro 1998

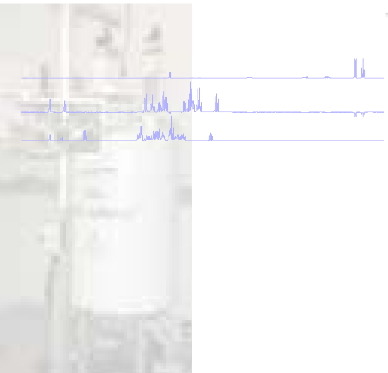


Electronic nose
Skov & Bro 2004



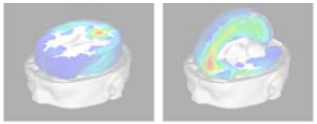
Other Examples

DOSY
Toft et al 2004

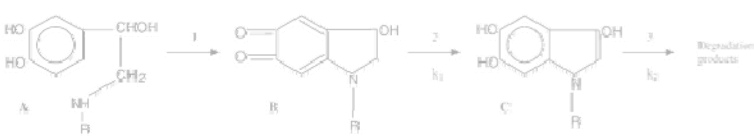


5-way analysis
Bro 1997

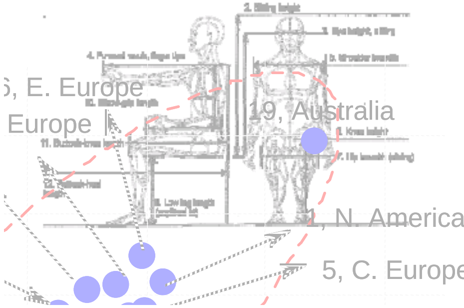
EEG
Miwakeichi et al 2004



Quantify catecholamine in urine

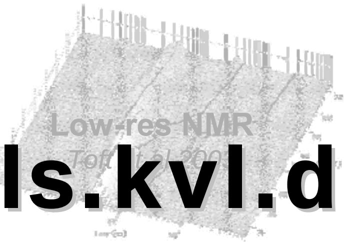


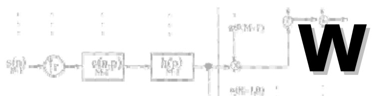
Anthropometry
S Lee 2006



Light-induced oxidation of cheese
Andersen & Bro 2004

Low-res NMR





Time-intensity
Ovejero-López et al, 2004

www.models.kvl.dk