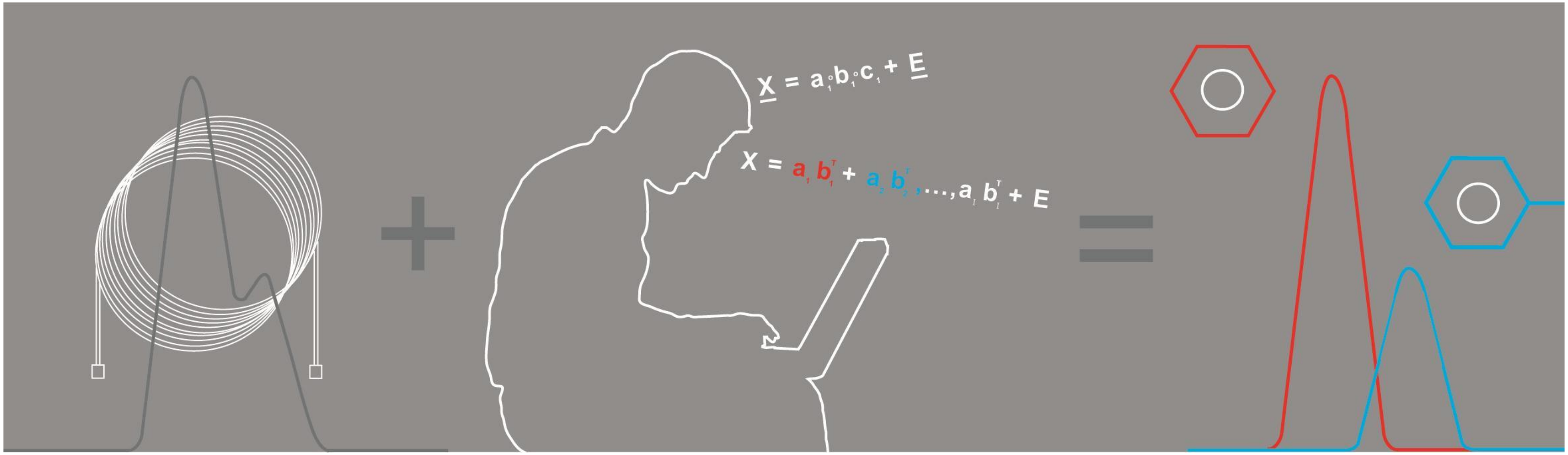
UNIVERSITY OF COPENHAGEN

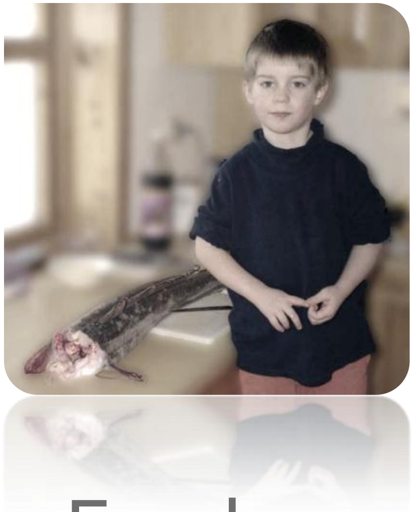
Faculty of Life Sciences

Practical Problems in Tensor Modeling

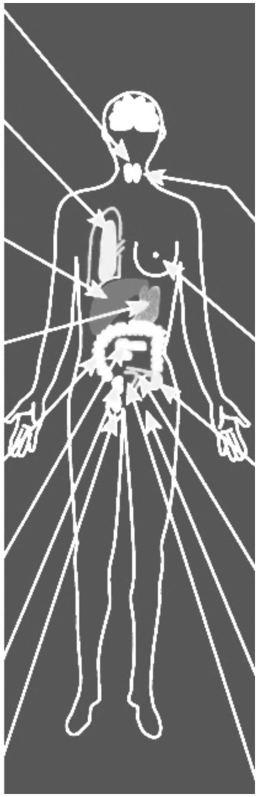
(In chemometrics)

Rasmus Bro


$\underline{X} = a_1 \circ b_1 \circ c_1 + \underline{E}$
$X = a_1 b_1^T + a_2 b_2^T ,\ldots, a_I b_I^T + E$



Dioxin,
Environment,
Dose-response



Genomics,
Systems biology,
Cancer,
Diabetes,
Pharma
…

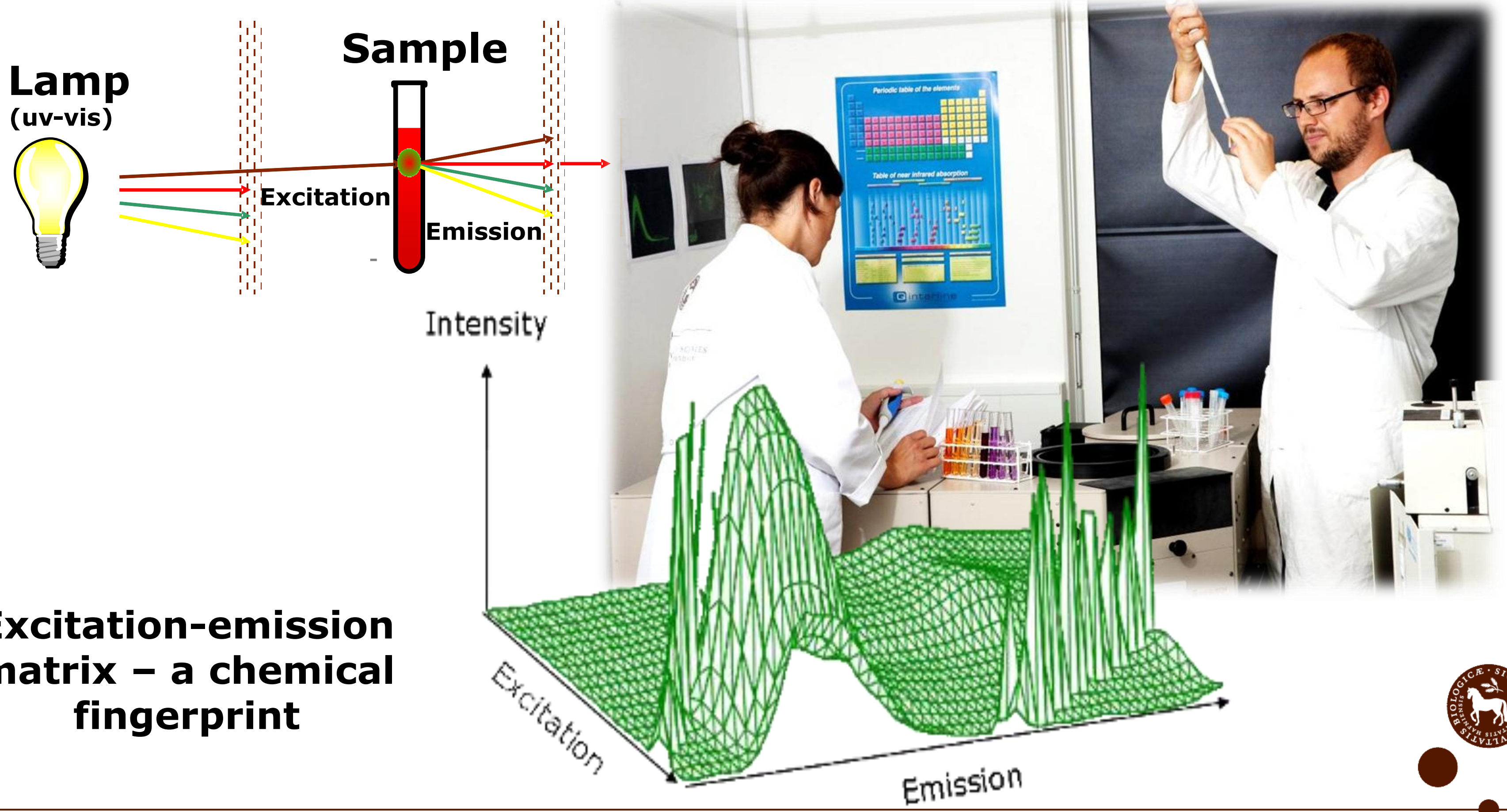
Food quality,
Raw material influence,
Production optimization

What we work with

Fluorescence
High resolution NMR
Mass spectromety
Near-infrared
Raman
Ultrasound
Hyperspectral Imaging
Chromatography
Imaging
…

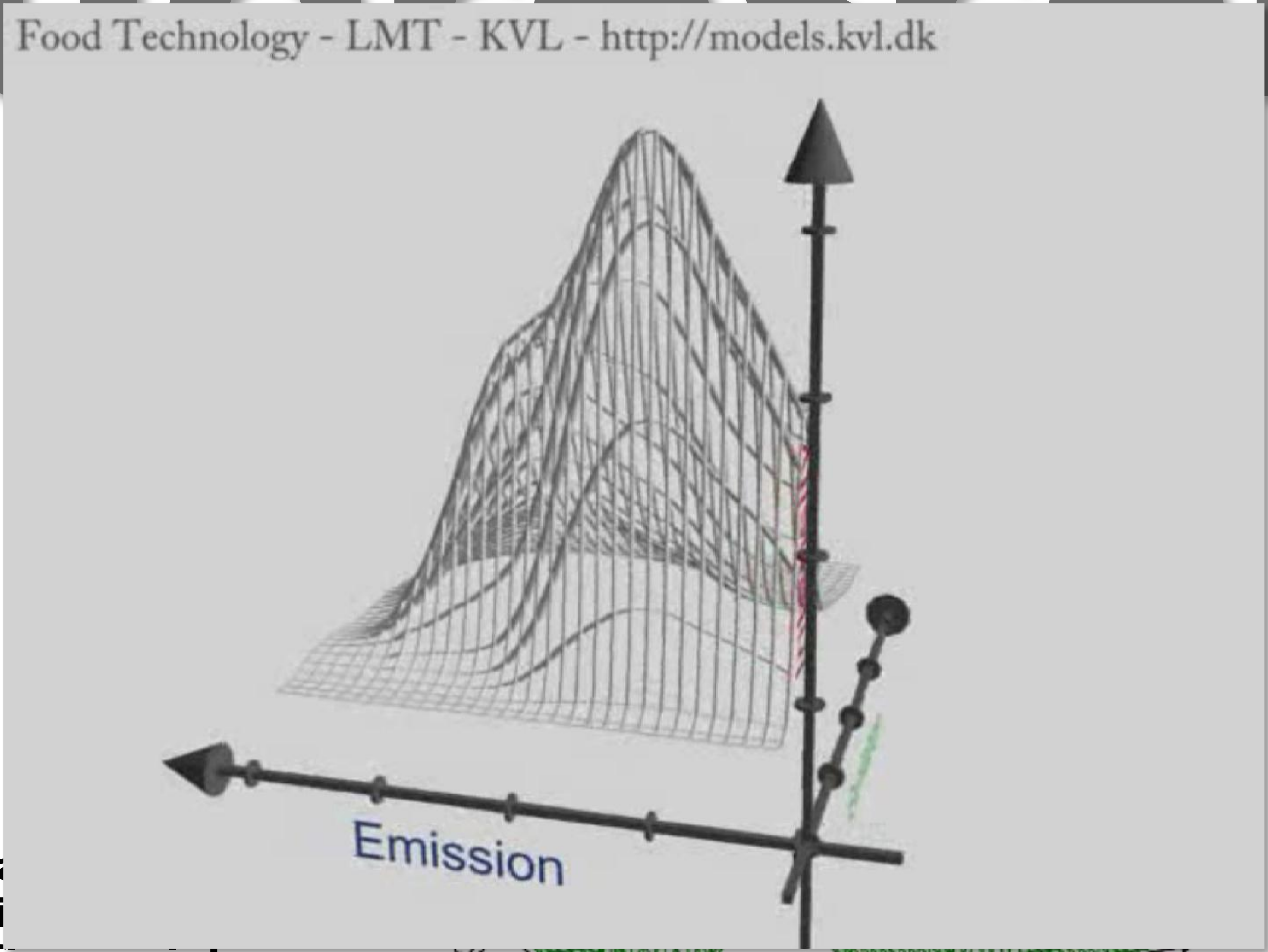
Data

Fluorescenc



Excitation-emission matrix – a chemical fingerprint

fluorescenc


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

Excita
matri
fingerprint

Basic problems

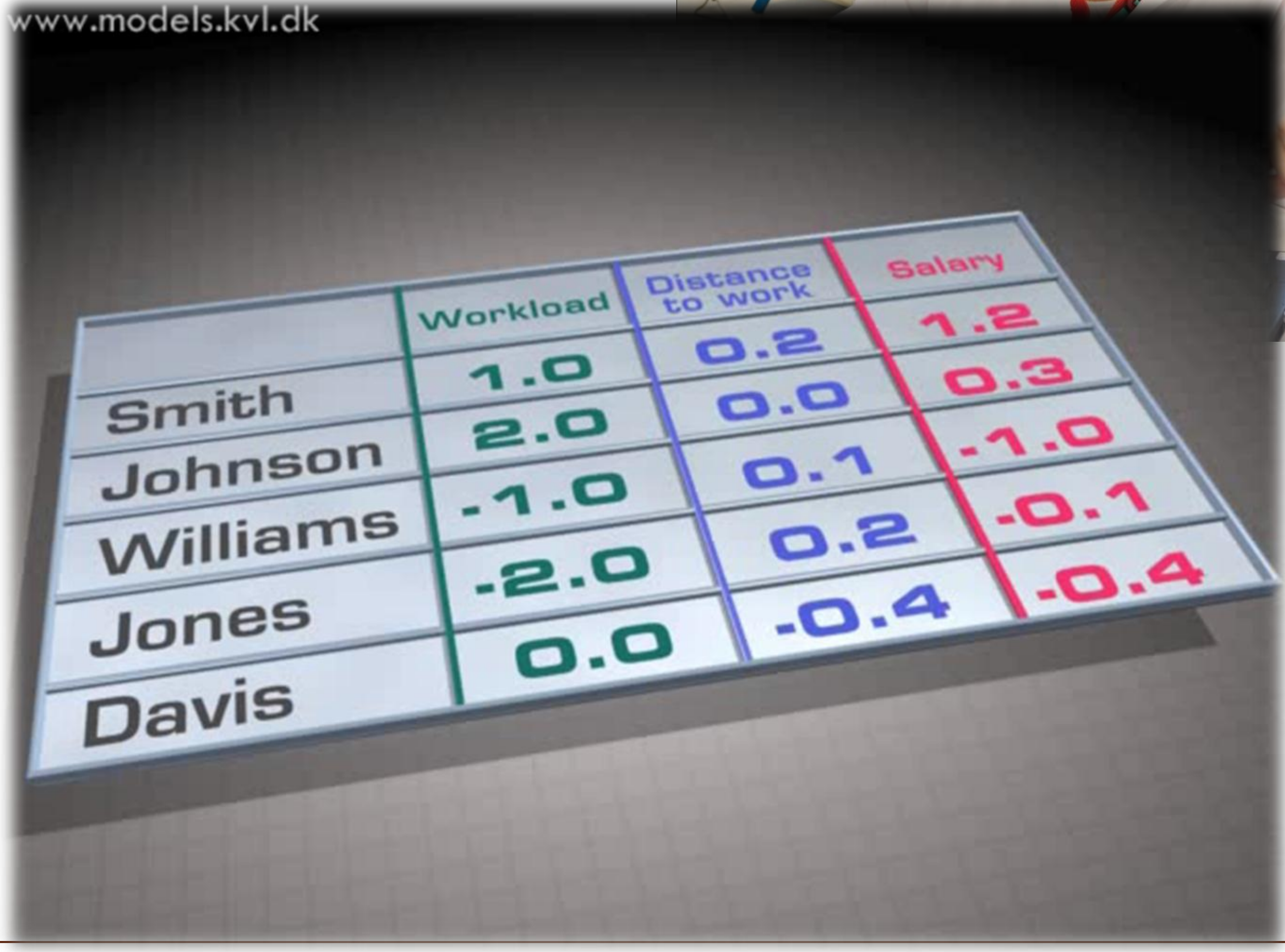
Plotting

Uncertainty estimates

Automated analysis

Interpretation

How to interpret
a scatter plot



www.models.kvl.dk

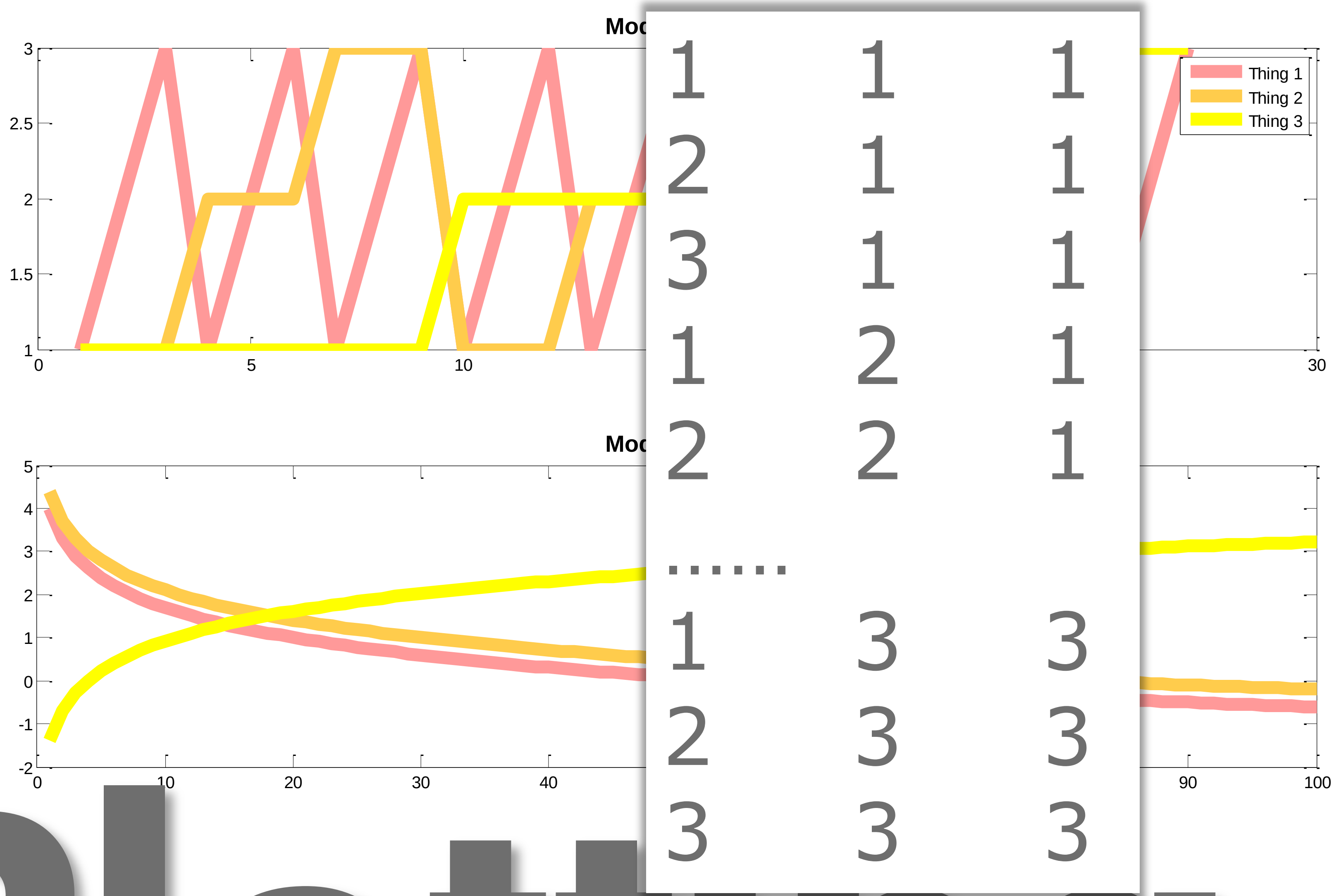
	Workload	Distance to work	Salary
Smith	1.0	0.2	1.2
Johnson	2.0	0.0	0.3
Williams	-1.0	0.1	-1.0
Jones	-2.0	0.2	-0.1
Davis	0.0	-0.4	-0.4

Plotting

- Two-way PCA - orthonormal basis (loadings)
- Hence distances in scores reflect both manifest and latent distances

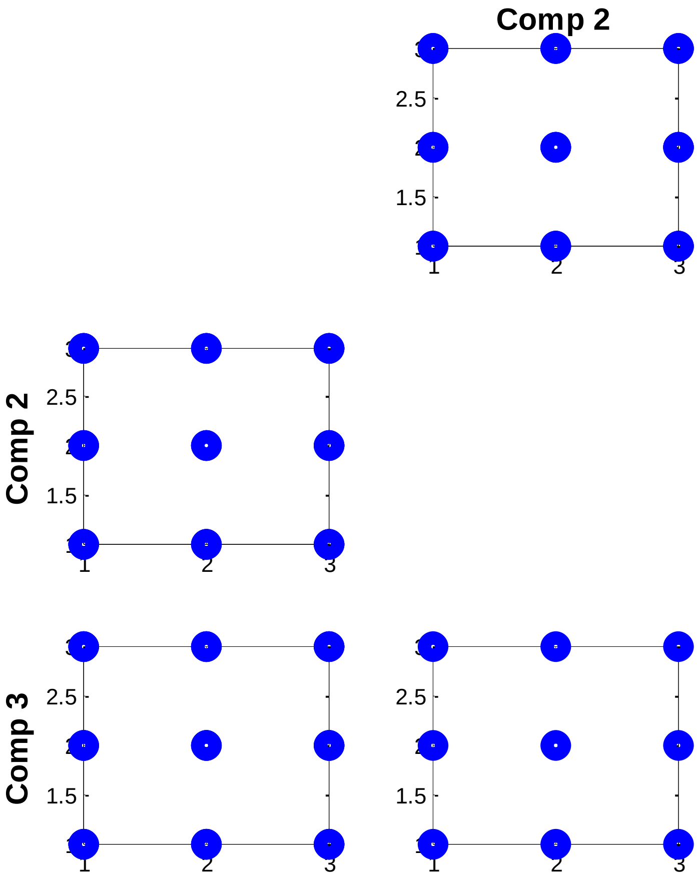
- PARAFAC/Tucker - Oblique bases
- Distances reflect only latent distances not manifest

Plotting

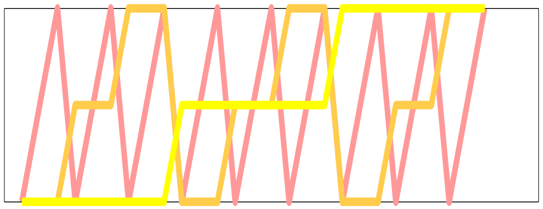


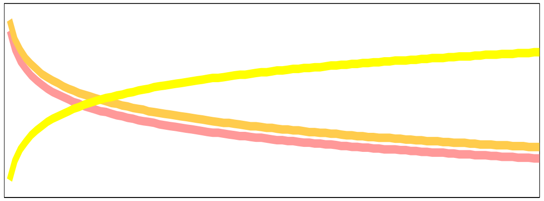
1	1	1
2	1	1
3	1	1
1	2	1
2	2	1
......		
1	3	3
2	3	3
3	3	3

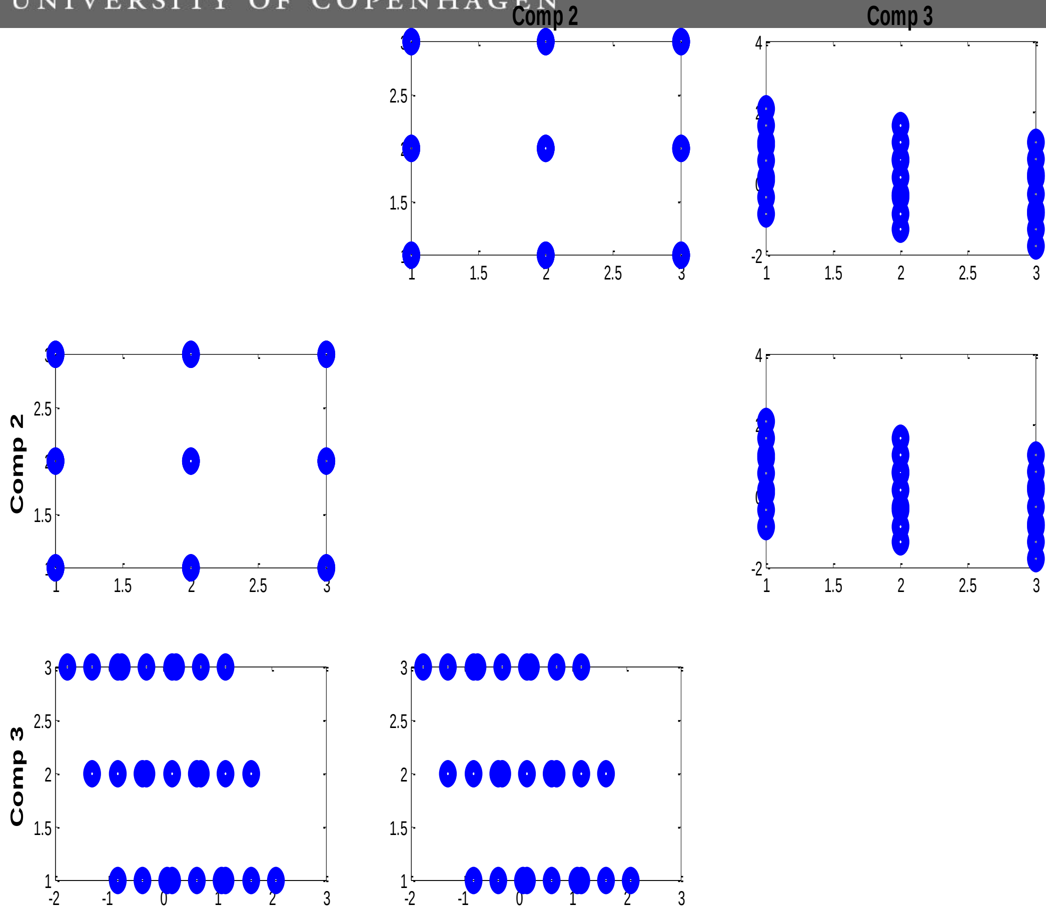
Plotting



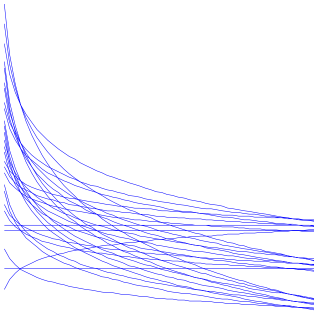
Plotting the first mode component in scatter plots **reflecting** ***latent*** **variation**

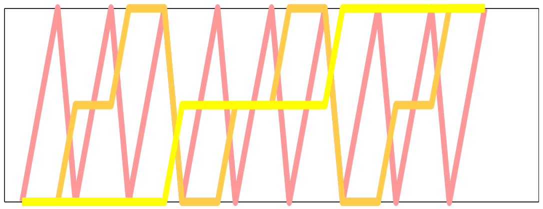


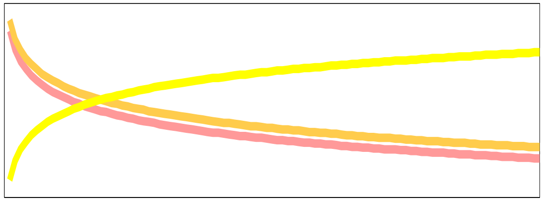




Plotting on an **orthogonal basis reflecting raw data distances**



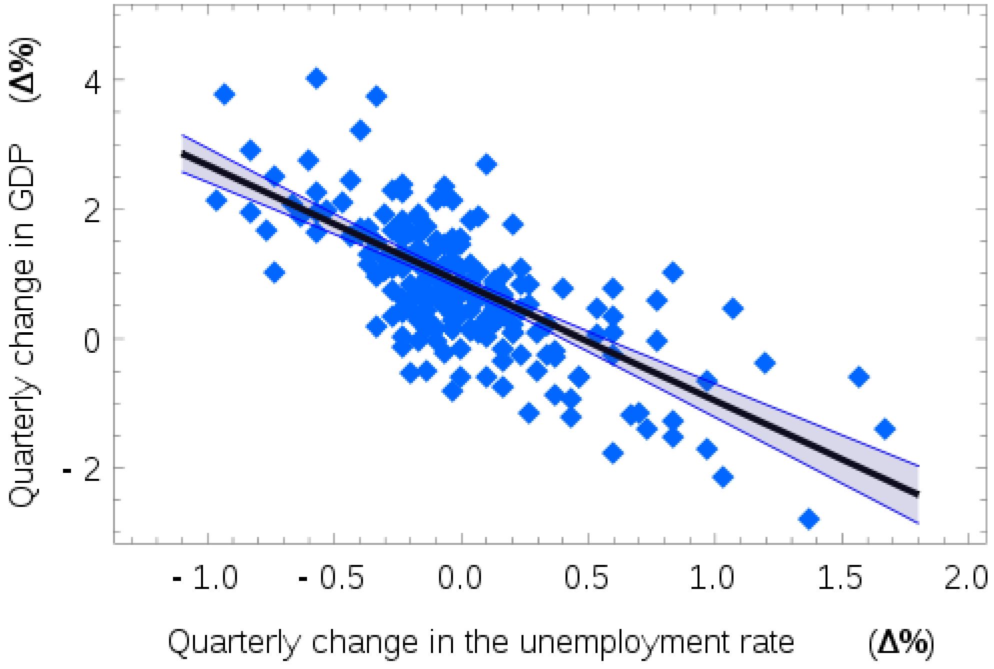




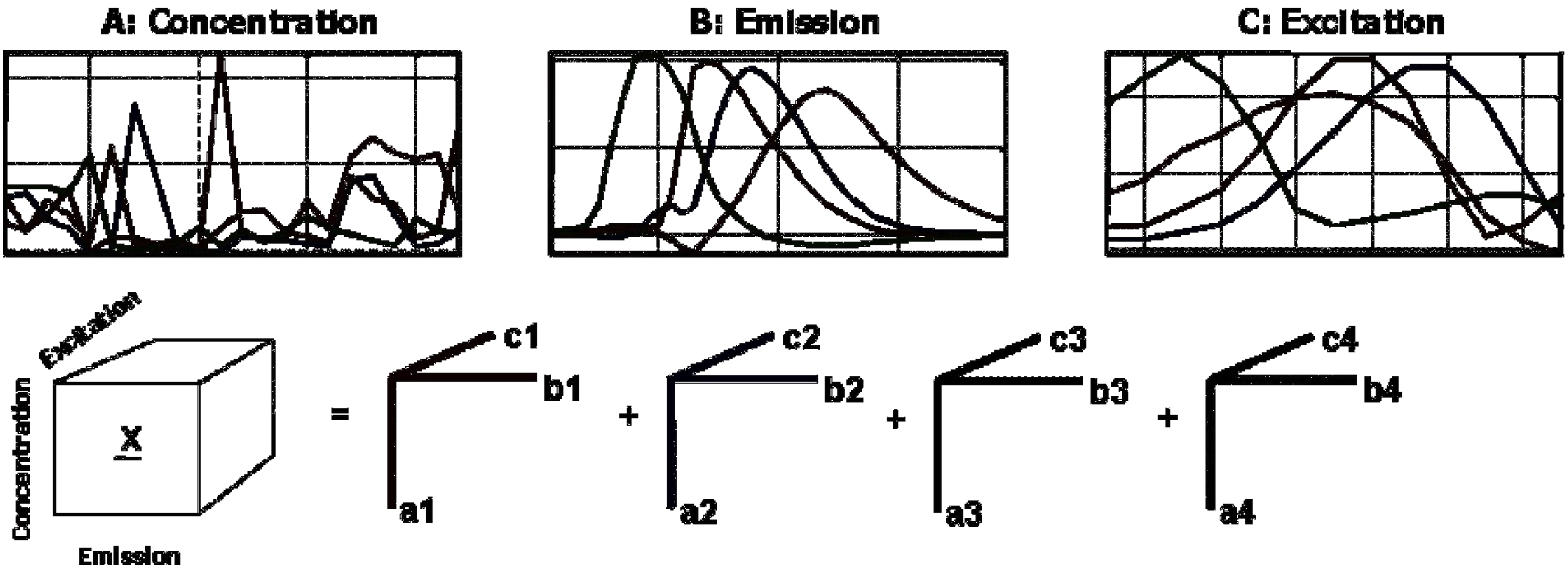
Uncertainty of parameters

$$S^2 = \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2}{N-1}$$

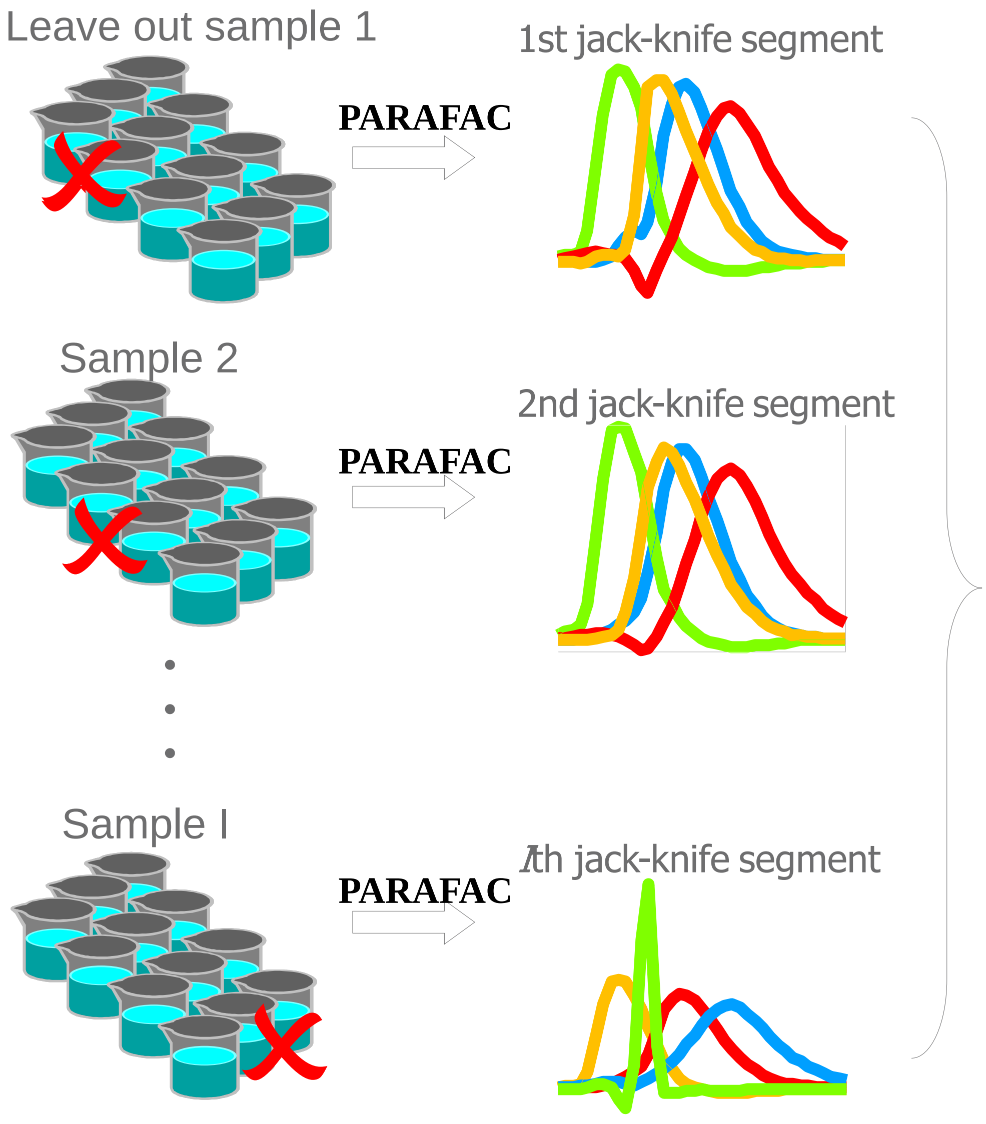
No degrees of freedom in PARAFAC (and probably not in other multilinear models)



PARAFAC on fluorescen



Jack-knifing the model



Leave out sample 1

PARAFAC

1st jack-knife segment

Sample 2

PARAFAC

2nd jack-knife segment

Sample I

PARAFAC

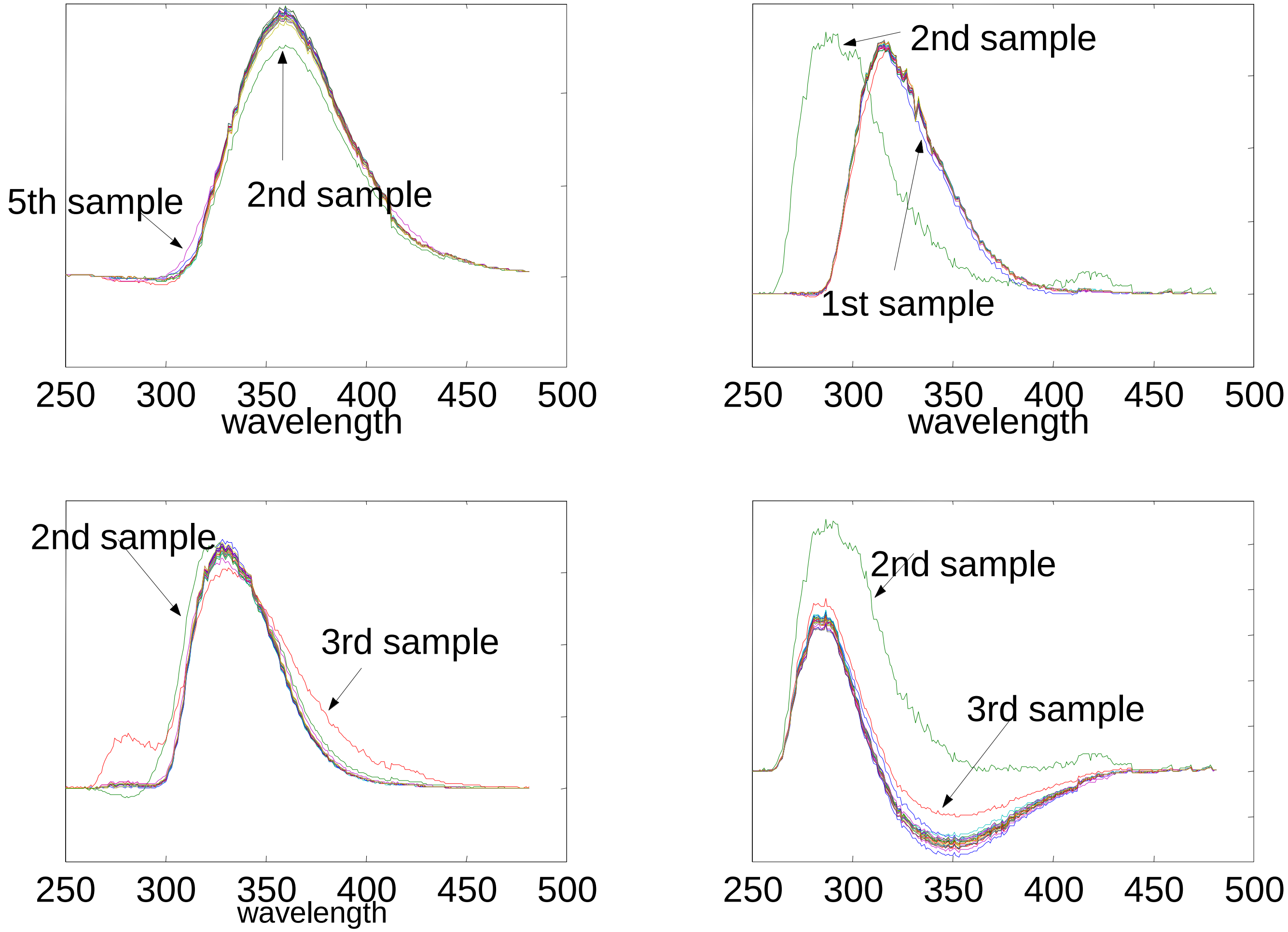
*I*th jack-knife segment

I PARAFAC sub-models:

- Standard error
- Outlier detection

J. Riu and R. Bro. Jack-knife technique for outlier detection and estimation of standard errors in PARAFAC models. Chemom. Intell. Lab. Syst. 65 (1):35-49, 2003.

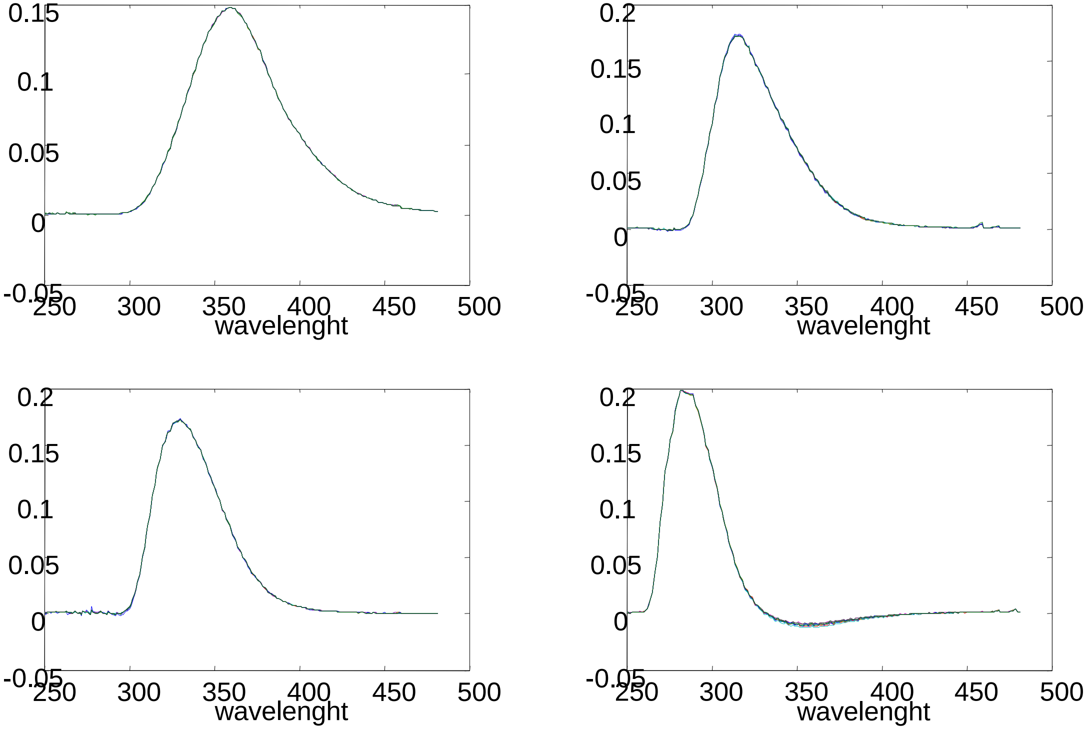
Jack-knifing the model



Emission spectral profiles

Jack-knifing the model

Removing low excitation and sample #2,3,5,10



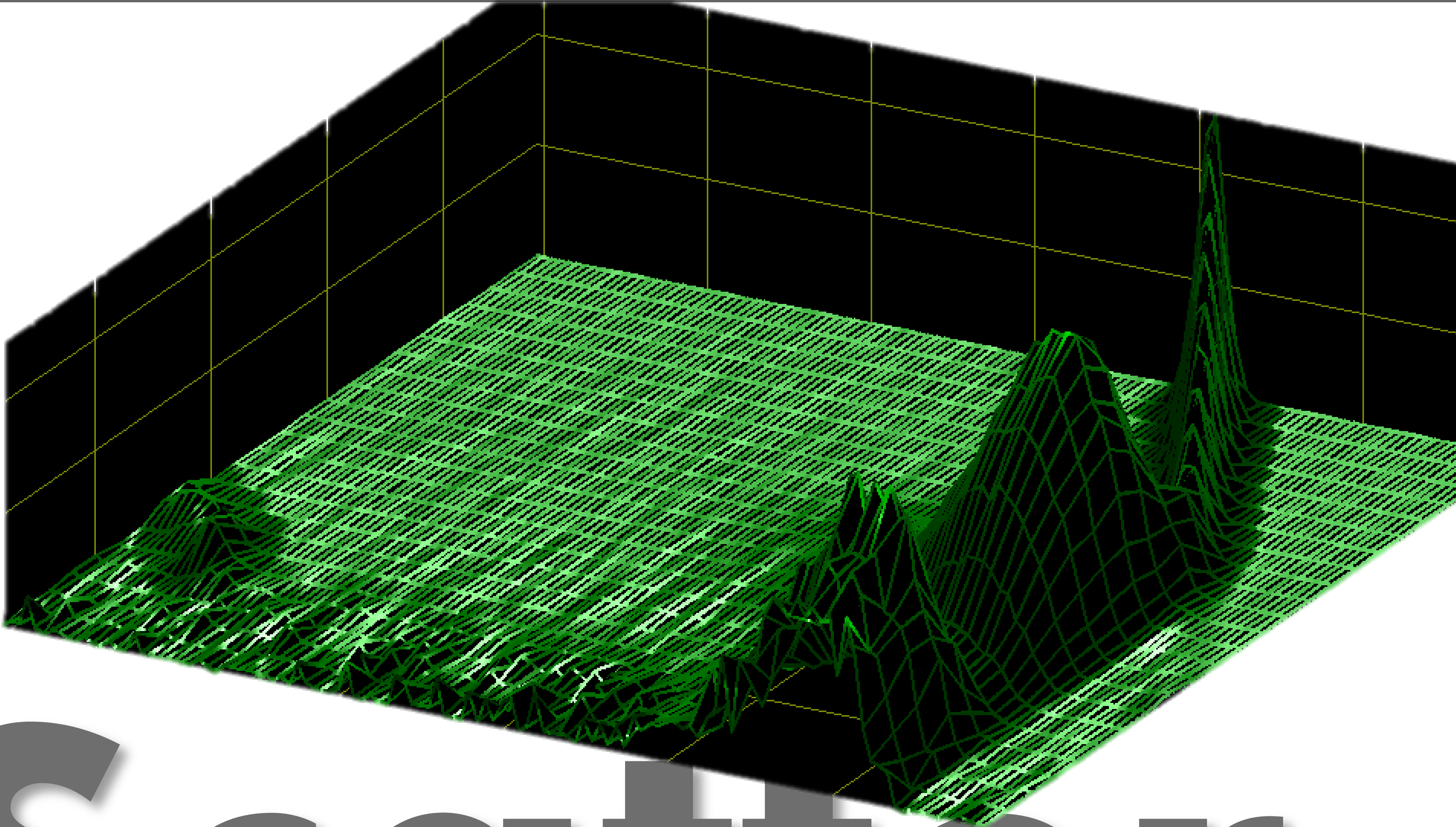
Emission spectral profiles

Automatic?

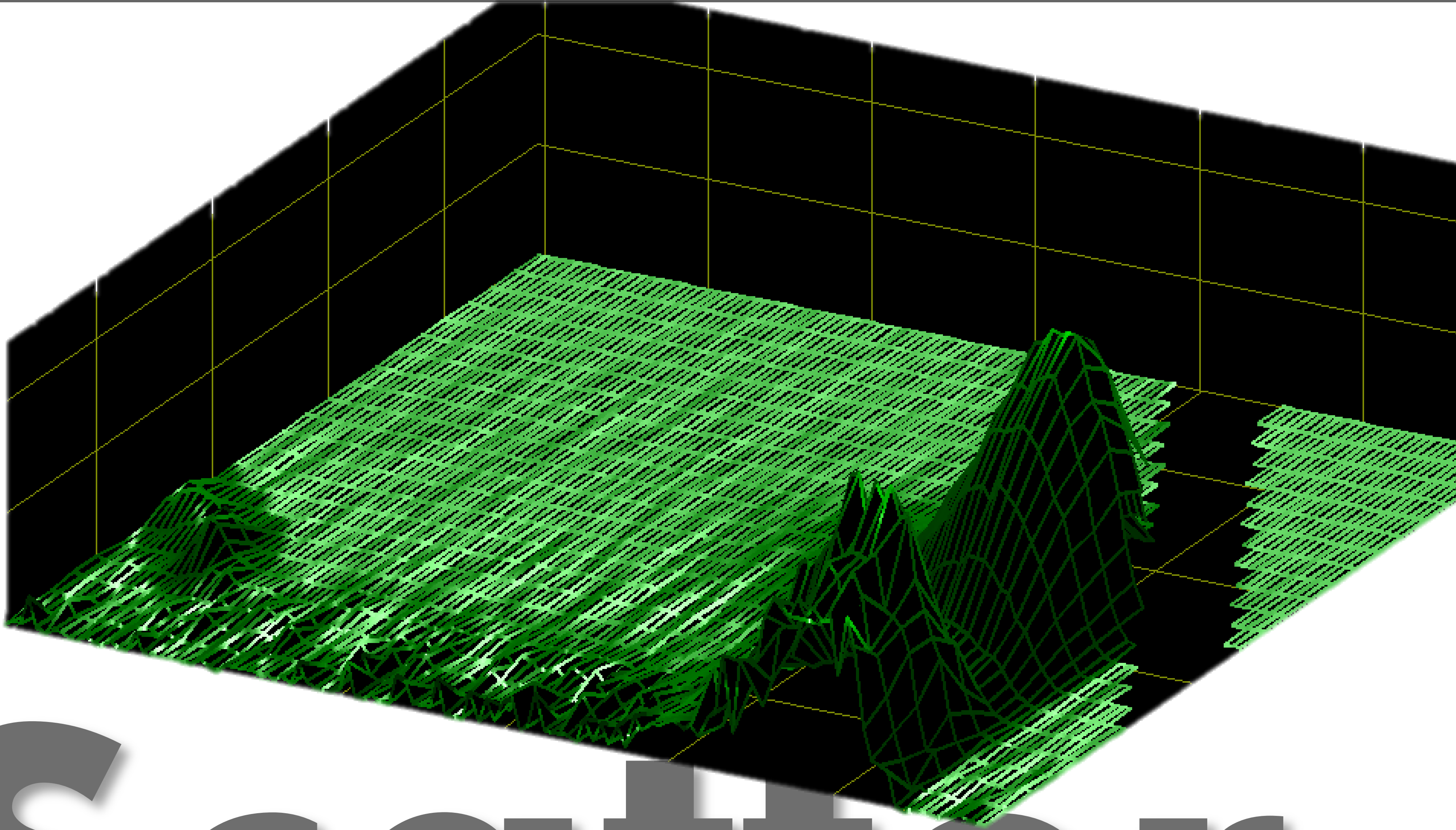
Meta-parameters

Goodness

Result



Scatter



Scatter

Long story …

Must be approximately valid

- Sufficient number of adequate samples
- Sufficient spectral resolution
- Beers law valid

Then decide

- Low excitation wavelengths to exclude
- How to handle Rayleigh scattering
- Number of components to use
- Outliers to exclude

EEMizer

Goodness criterion

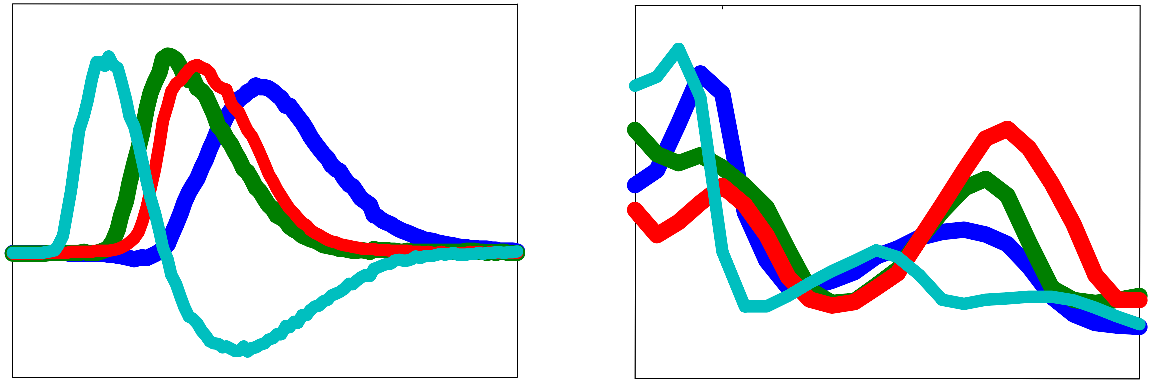
Goodness = Fit*CoreConsistency*Splithalf

$$FIT = 1 - \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} e_{ijk}^{2}}{\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} x_{ijk}^{2}}$$

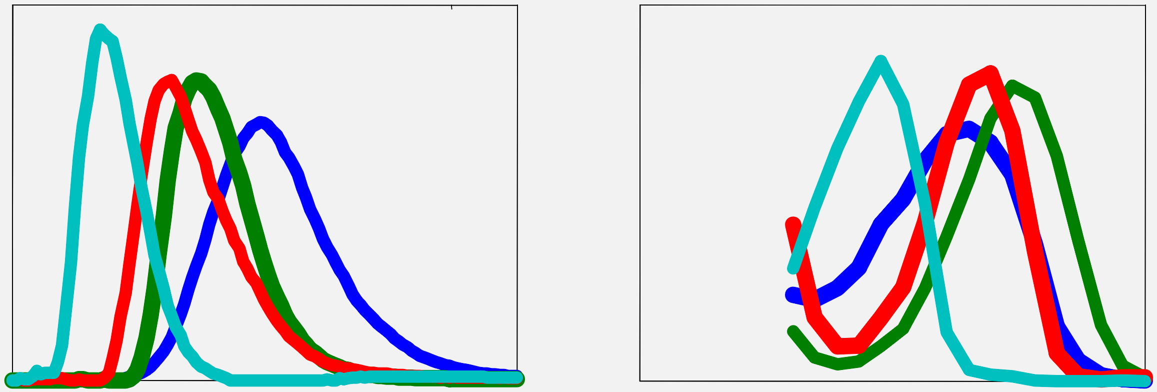
$$COREC = 100\left(1 - \frac{\sum_{d=1}^{F}\sum_{e=1}^{F}\sum_{f=1}^{F}\left(g_{def} - t_{def}\right)^{2}}{F}\right)$$

EEMizer

Before EEMizer

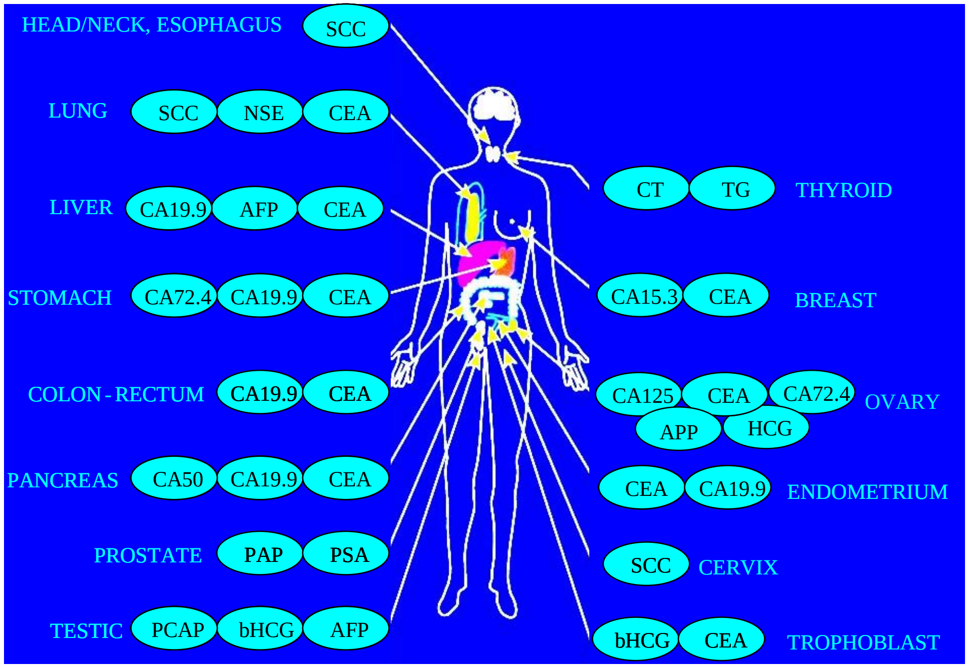


After EEMizer



EEMizer result

Traditional approach for cancer diagnostics and monitoring: Biomarkers



HEAD/NECK, ESOPHAGUS: SCC

LUNG: SCC, NSE, CEA

LIVER: CA19.9, AFP, CEA

STOMACH: CA72.4, CA19.9, CEA

COLON - RECTUM: CA19.9, CEA

PANCREAS: CA50, CA19.9, CEA

PROSTATE: PAP, PSA

TESTIC: PCAP, bHCG, AFP

THYROID: CT, TG

BREAST: CA15.3, CEA

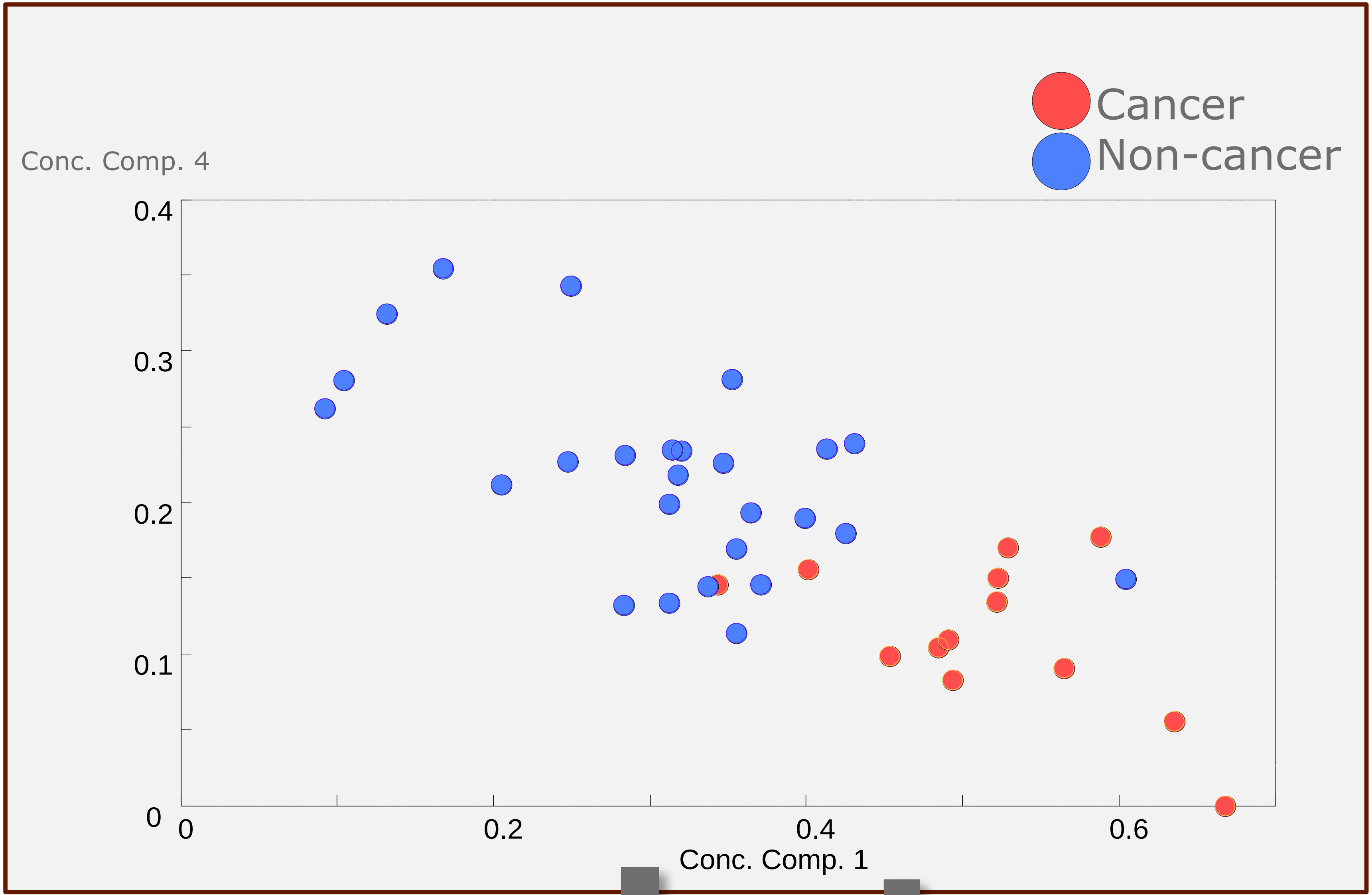
OVARY: CA125, CEA, CA72.4, APP, HCG

ENDOMETRIUM: CEA, CA19.9

CERVIX: SCC

TROPHOBLAST: bHCG, CEA

cancer



It works!

Conclusion

Still needed
Better algorithms
Better statistics
Better software

m-files, e-courses, data sets, etc.

www.models.life.ku.dk

If you want lots of papers on applied tensor analysis, come by with a USB