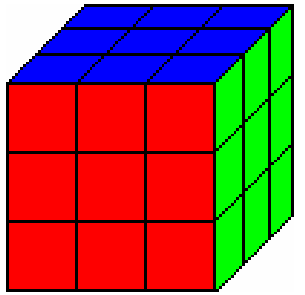
Multiway Data Analysis

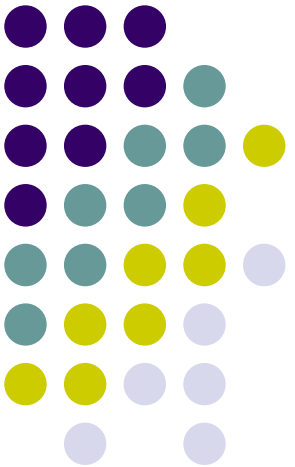


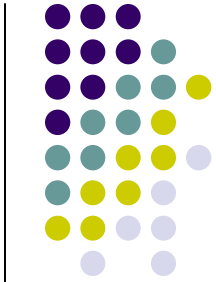
Johan Westerhuis

Biosystems Data Analysis

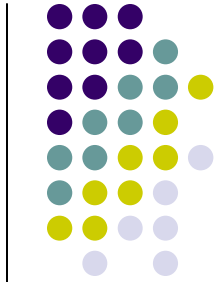
Swammerdam Institute for Life Sciences

Universiteit van Amsterdam

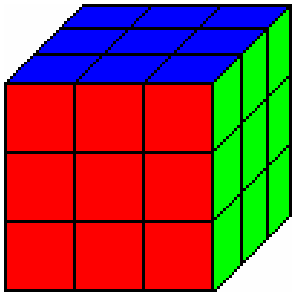




- Three-way Data
- Three-way Models
- Three-way Applications

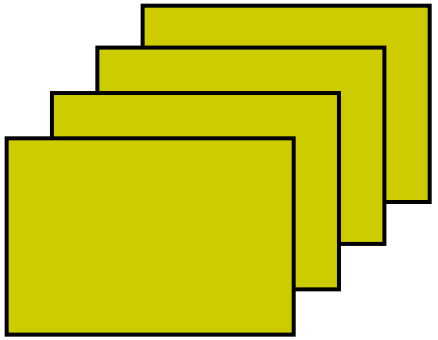


Three-way Data



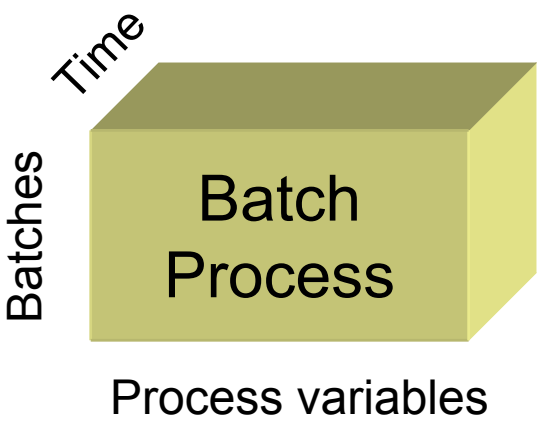
Three-way data

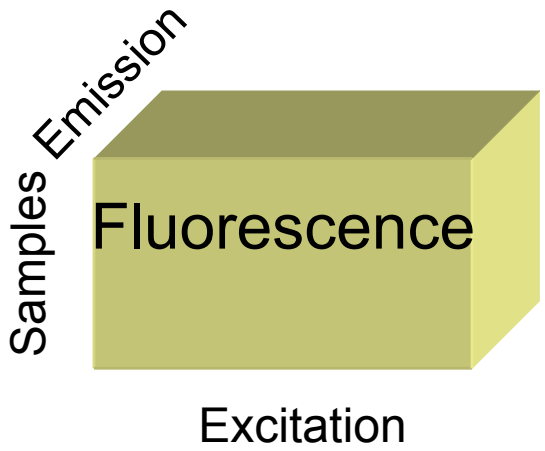
- Three-way data is a set of two-way matrices of the same objects and variables.

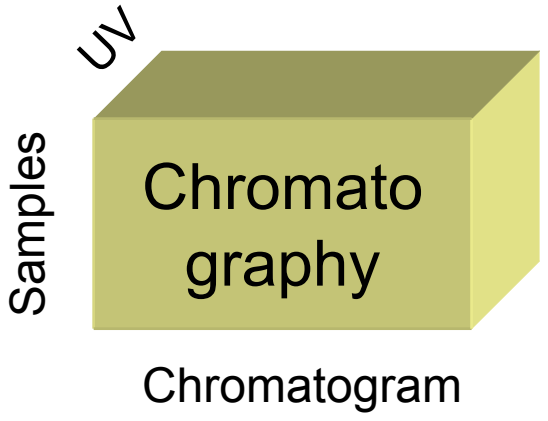


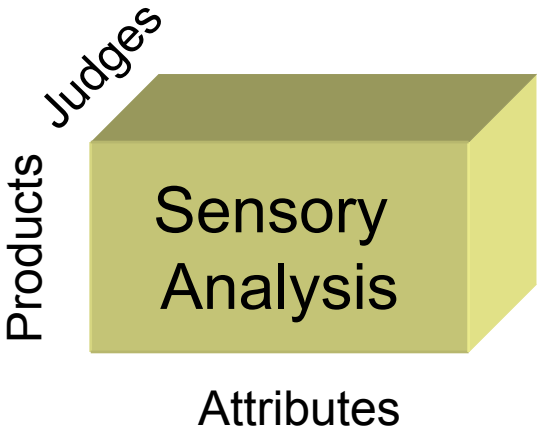
- IR, Raman, NMR spectra of the same samples will not give a three-way data set, but a multi-block data set.

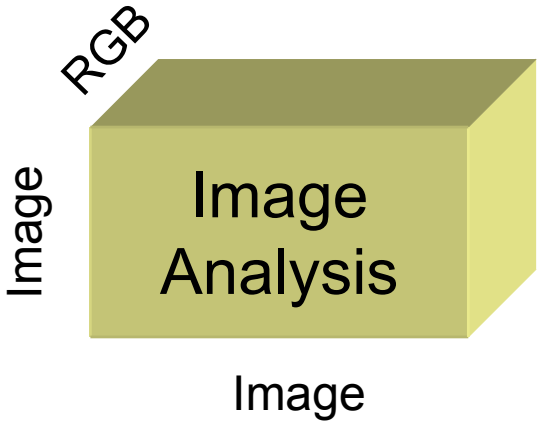
Examples of three-way data



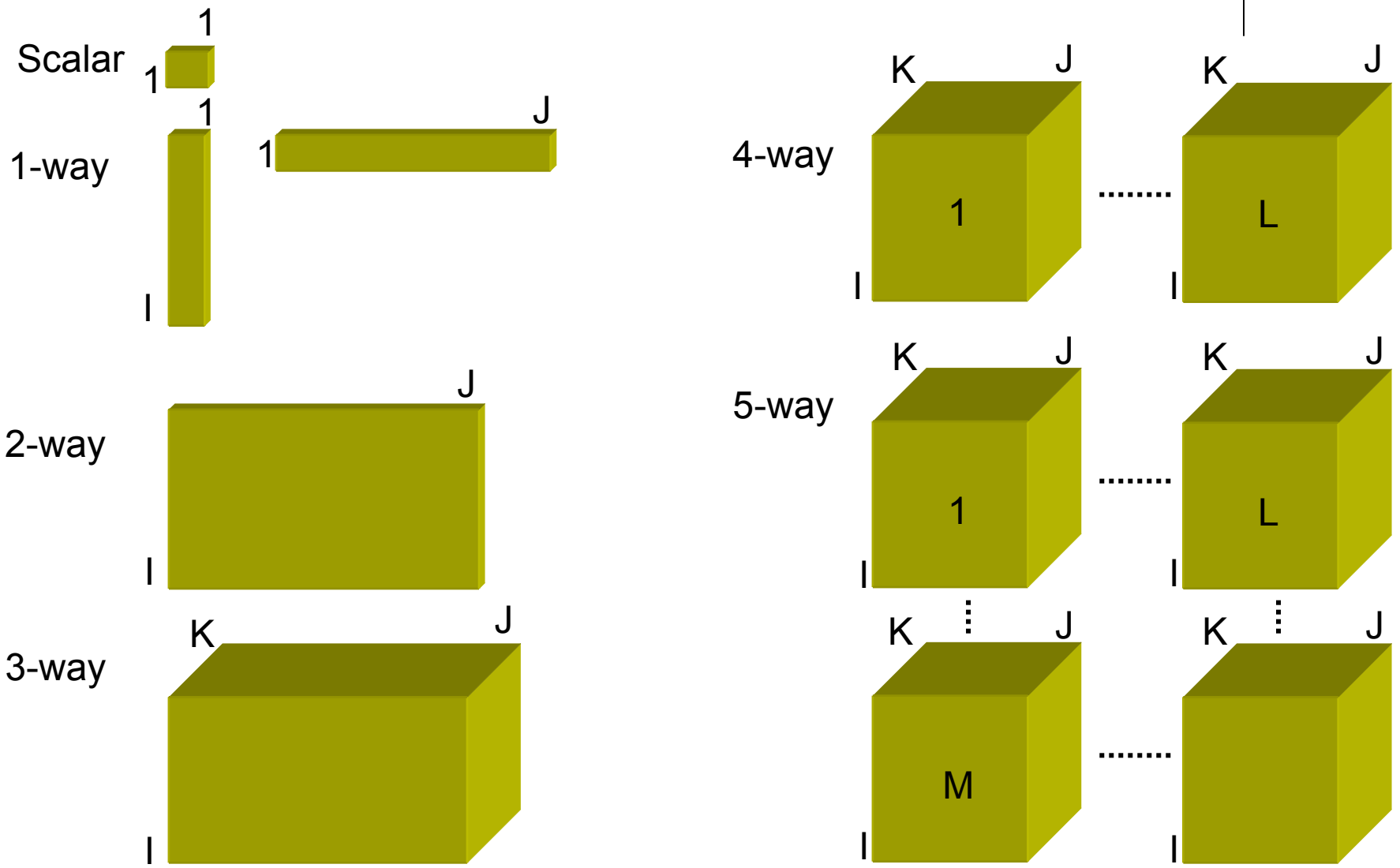




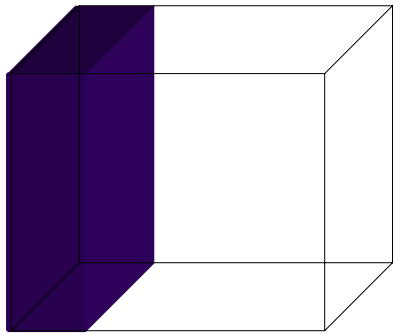




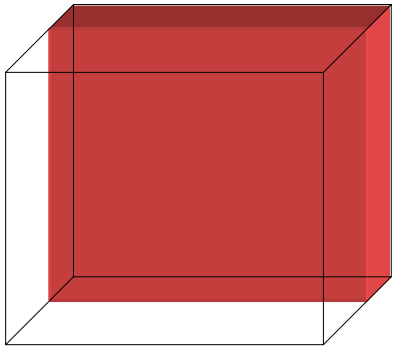
From noway to multi-way



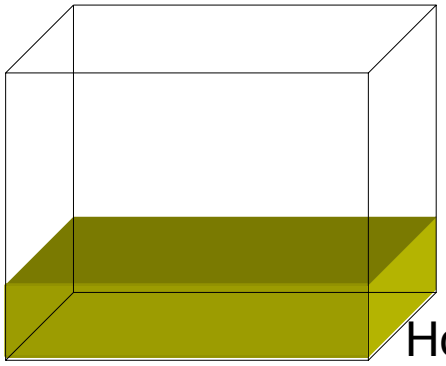
Slabs and tubes



Vertical slab

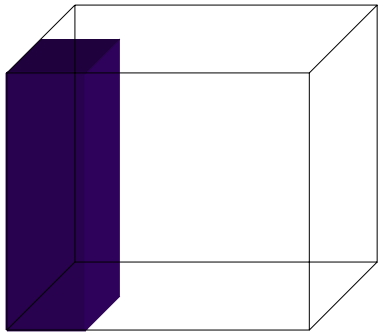


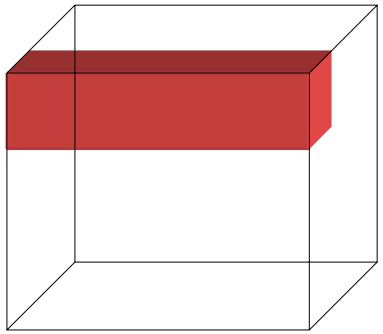
Frontal slab



Horizontal slab

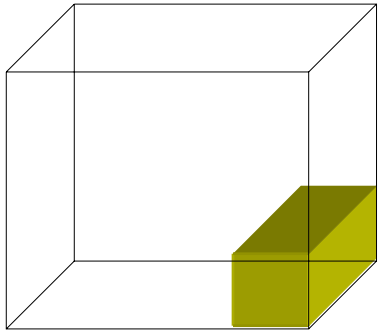
Vertical tube



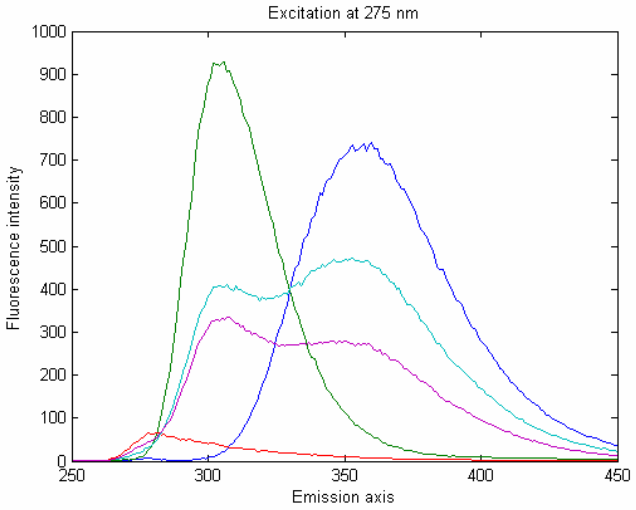


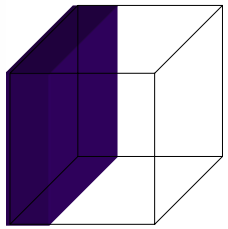
Horizontal tube

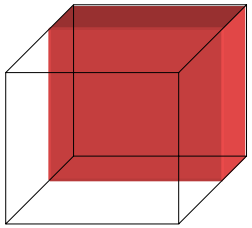
Lateral tube

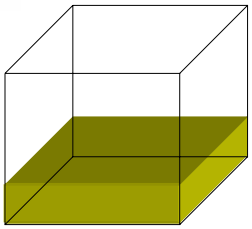


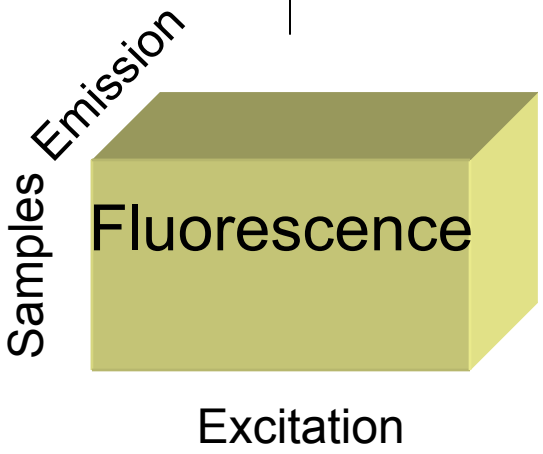
Three slabs of fluorescence data
5 Samples x 60 Excitation x 200 Emission

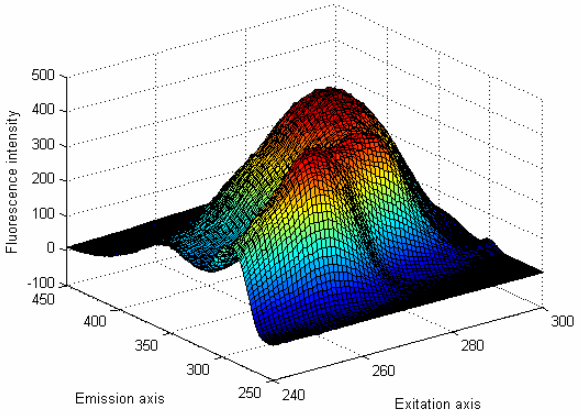


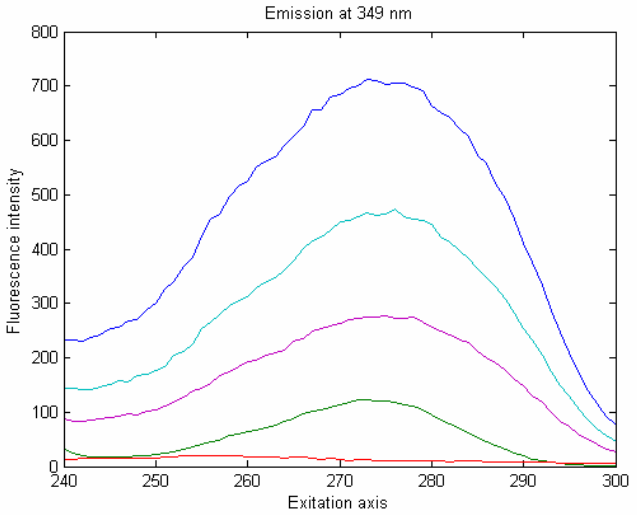






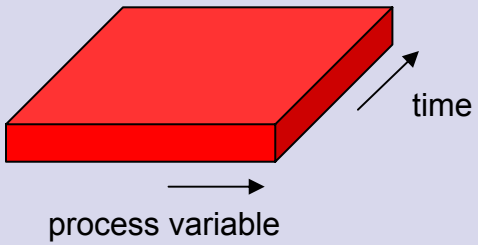




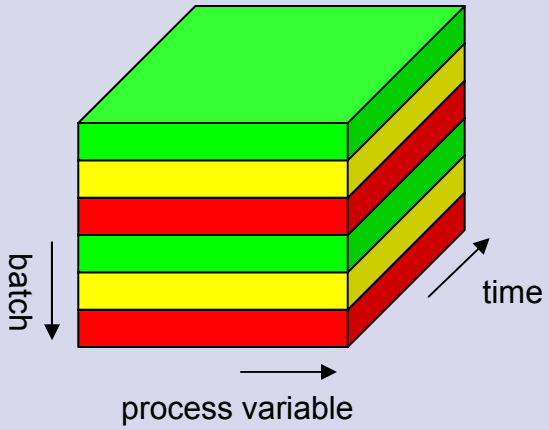


Three-way batch process data

- **'Engineering' process data *i.e.* temperature, pressure, flow rate**
- **Spectroscopic process data *i.e.* NIR, Raman, UV-Vis**

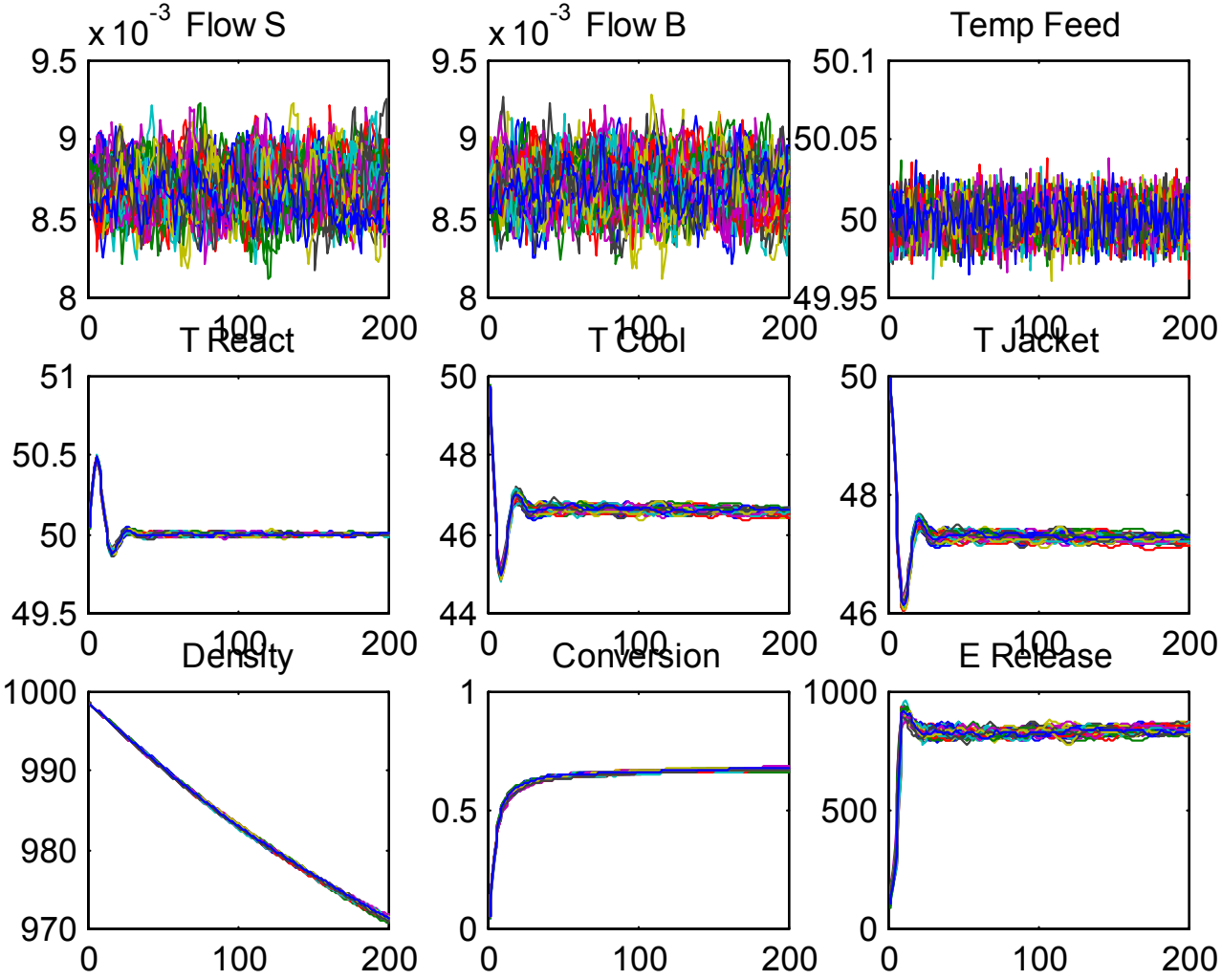


One batch
$\mathbf{X}$ $(J \times K)$

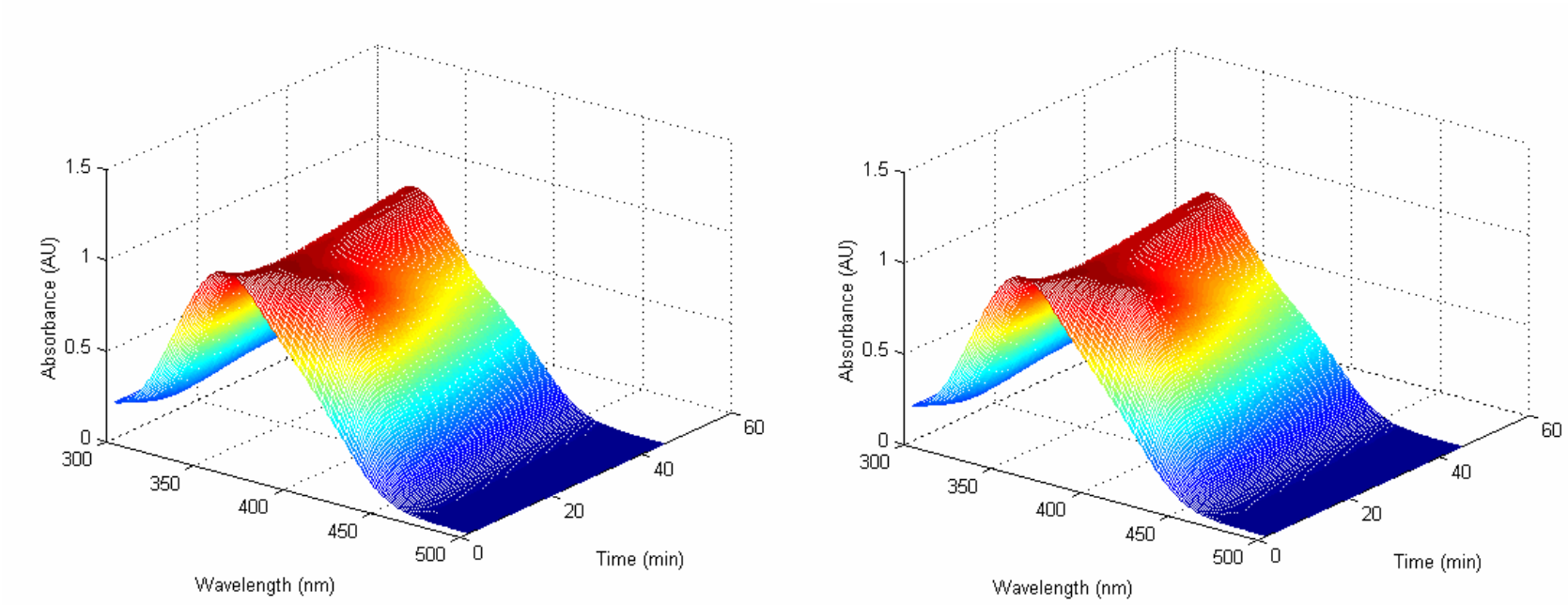


A series of batches
$\underline{\mathbf{X}}$ $(I \times J \times K)$

SBR batch process data
Engineering variables

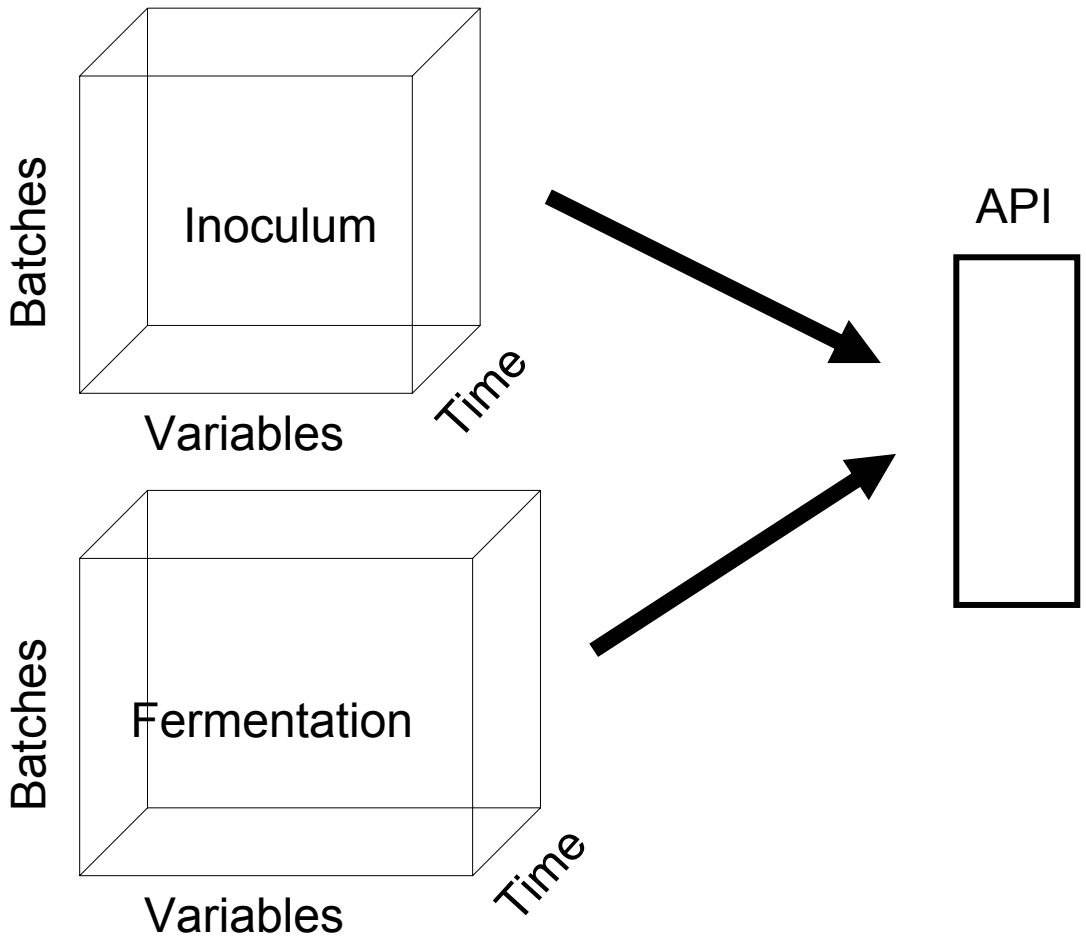


Spectroscopic three-way batch data

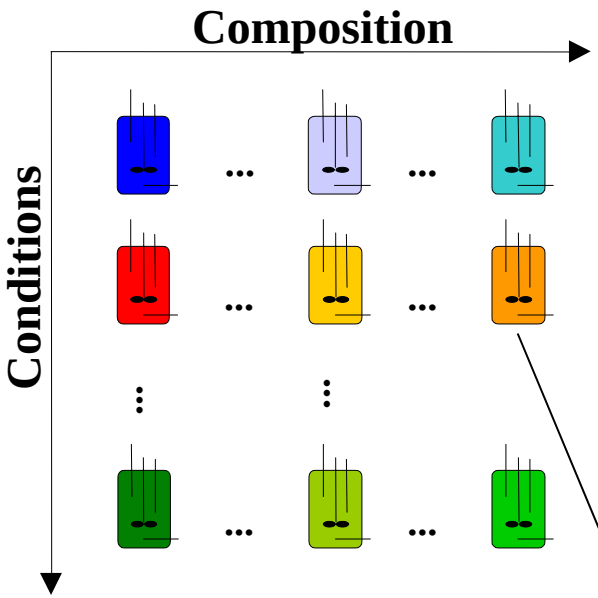


2 batch runs of a reaction followed with UV-Vis spectroscopy during 45 minutes

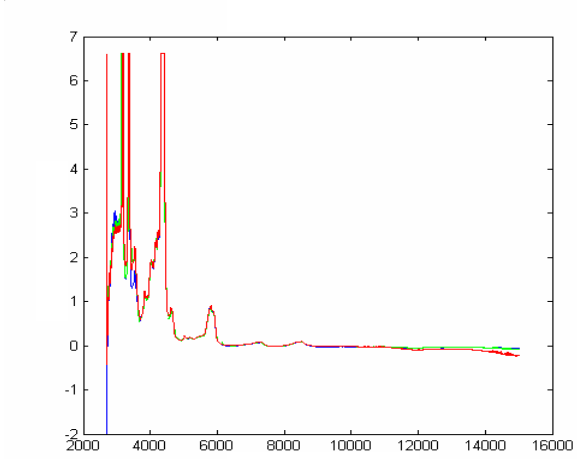
Batch Fermentation in two steps: Threeway multiblock

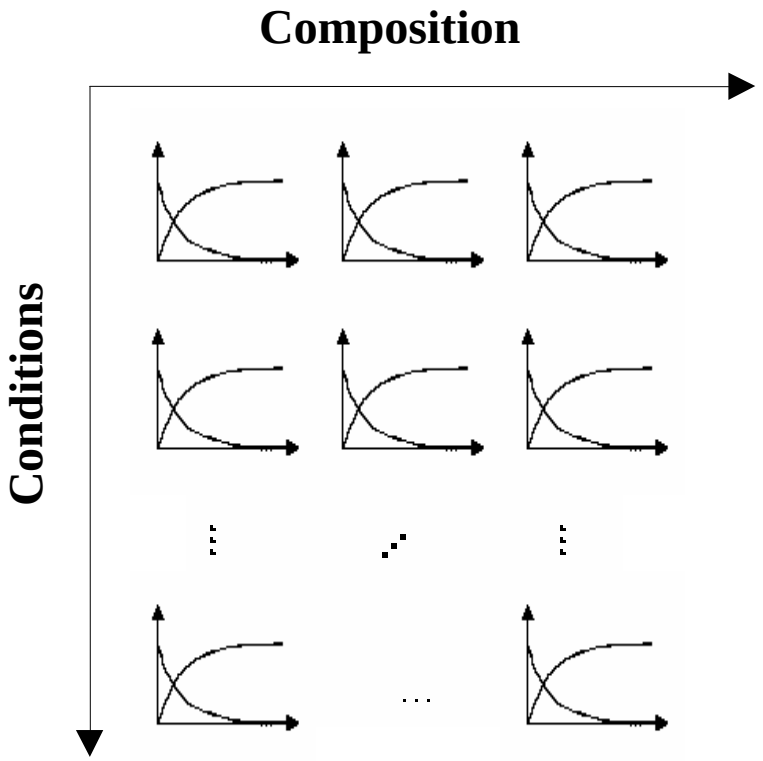


Four-way data in combinatorial catalysis



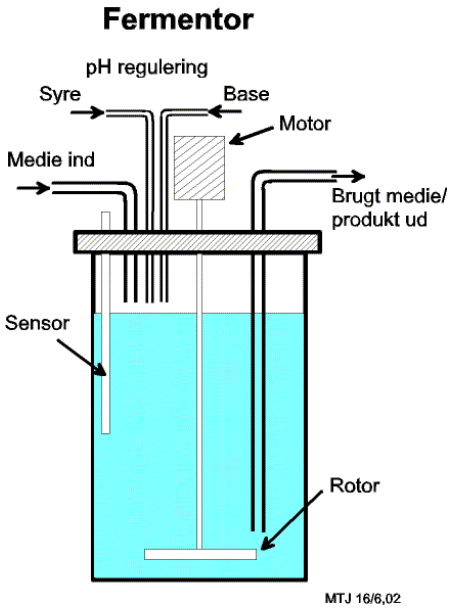
What we measure

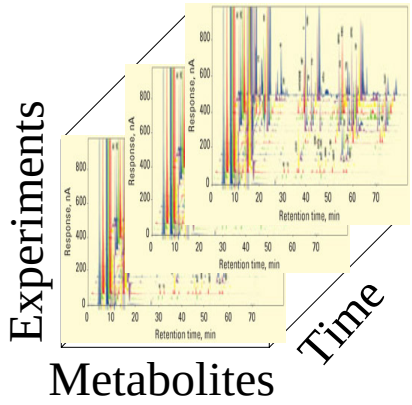


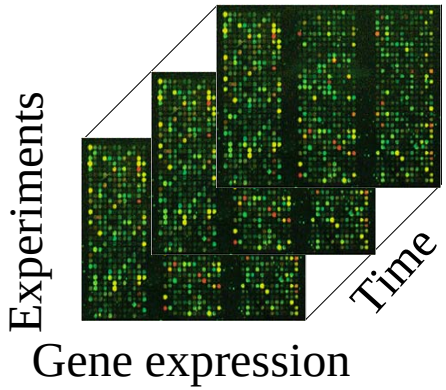


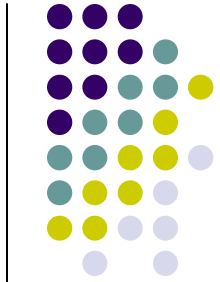
What we want

Multiway data from the Omics age

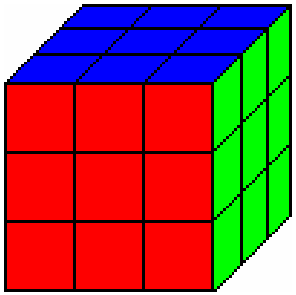




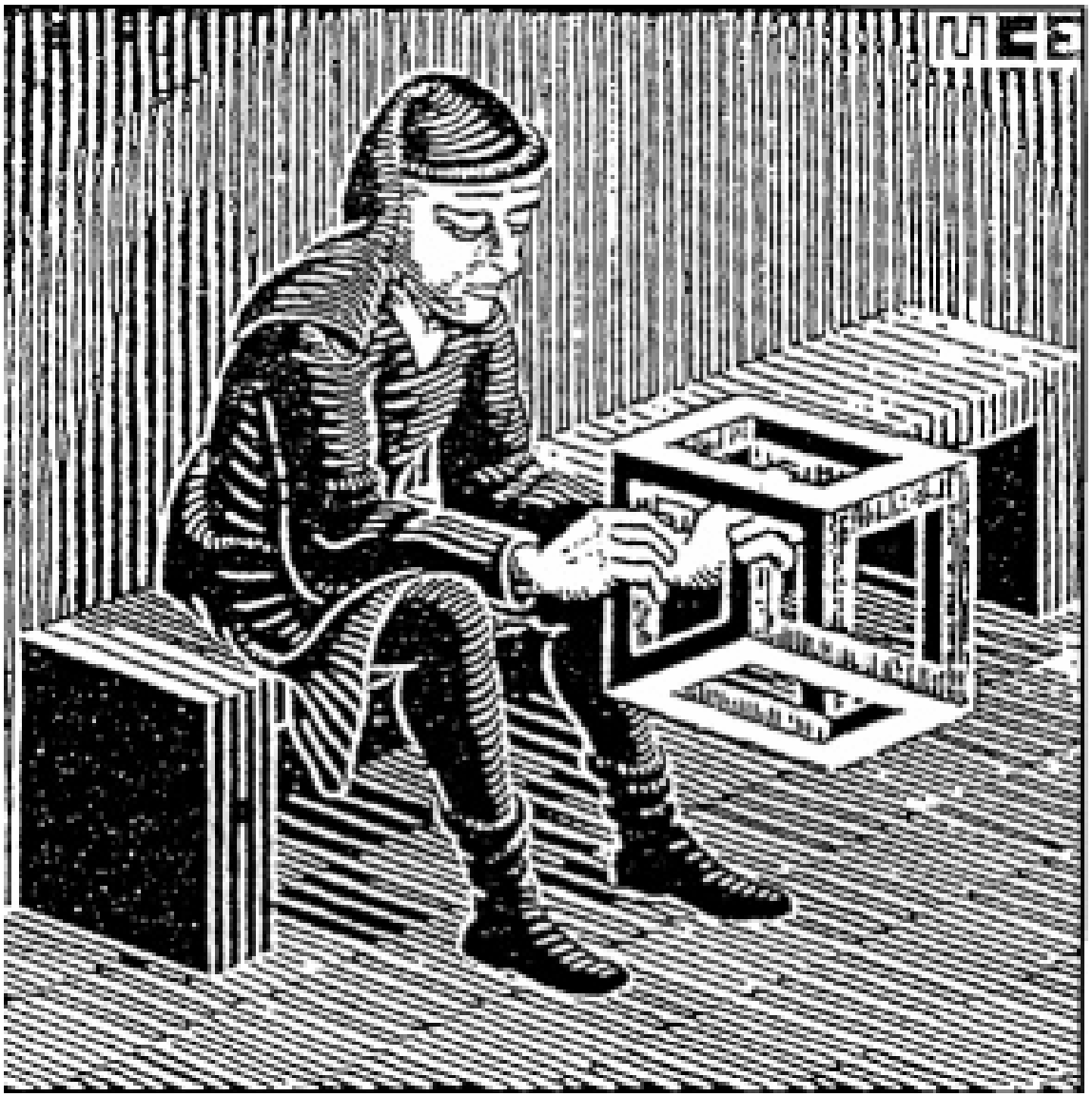




Three-way Models



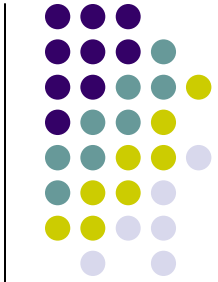
Some history



M.C. Escher:

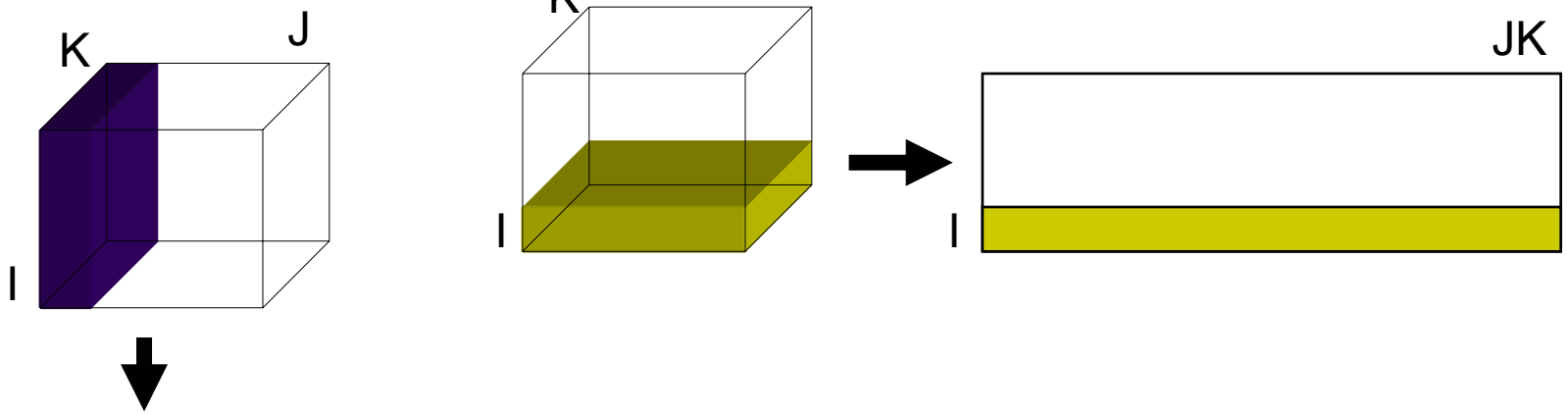
Small problem with orthogonality

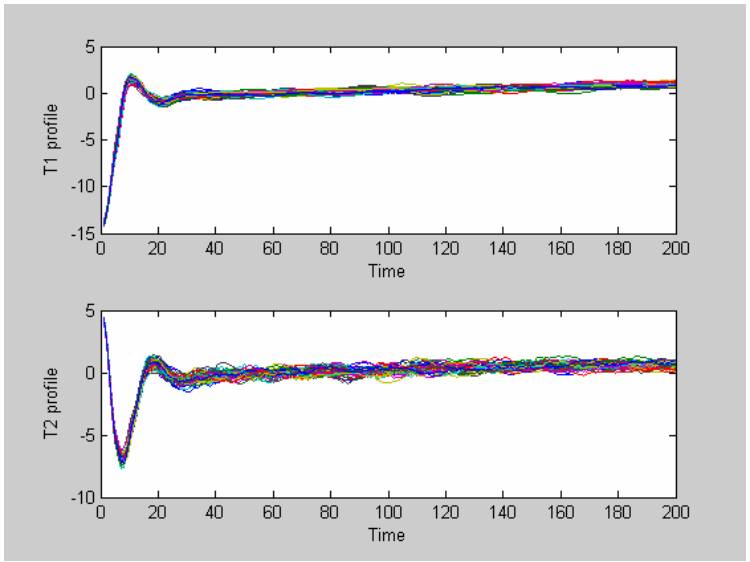
More history



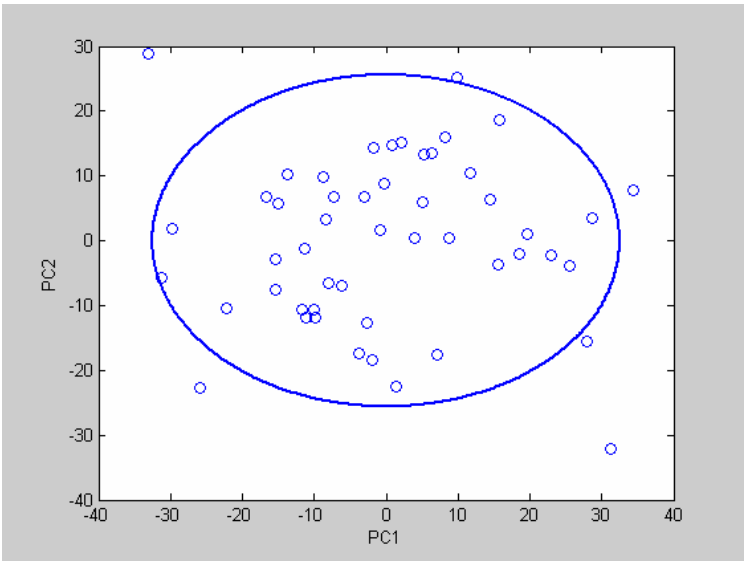
- Psychometrics (1944-1980)
  - Catell 1944: Parallel Proportional profiles (Common factors fitted simultaneously to many data matrices).
  - Tucker 1964: Tucker models
  - Carroll & Chang 1970: Canonical Decomposition (CANDECOMP)
  - Harshman 1970: Parallel Factor Analysis (PARAFAC)
- Chemistry
  - Ho 1978: Rank Annihilation (close to Parafac) on fluorescence data.
  - End 80's beginning 90's: Threeway methods to resolve LC-UV data.

Multiway PCA: Unfolding of three-way data





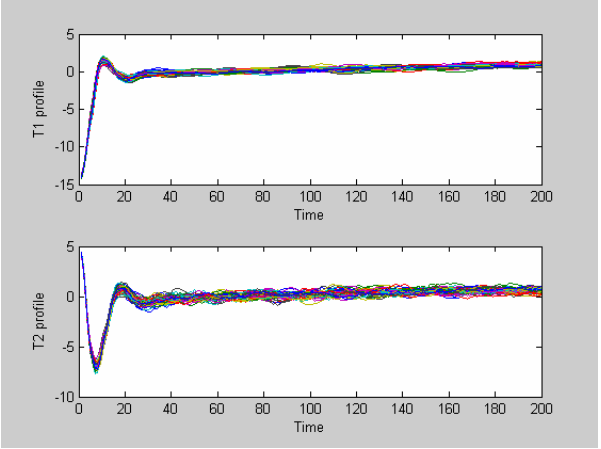
Wold



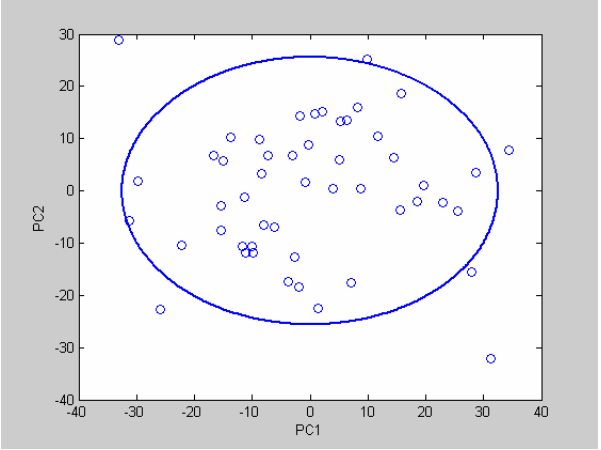
MacGregor

Two ways of unfolding
Different assumptions in MSPC

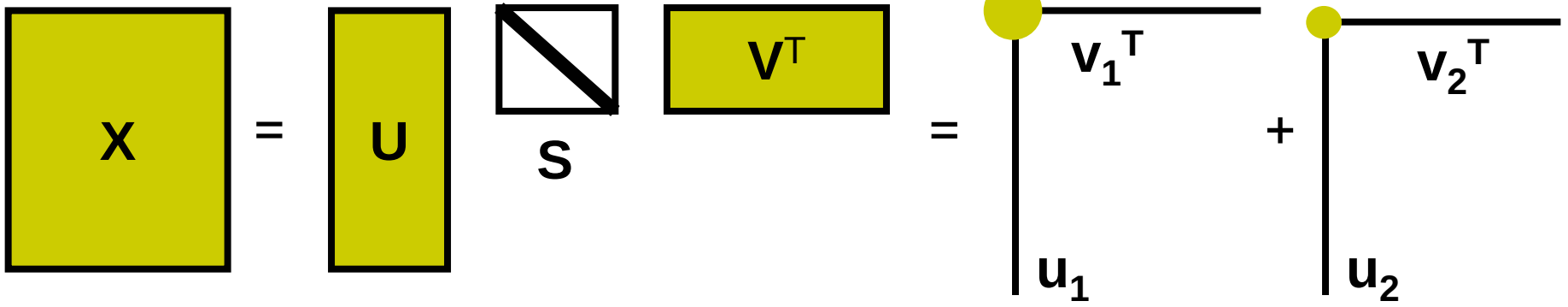
- Wold
  - Nonlinear behavior in the data
  - Batch trajectories are monitored
  - Online monitoring

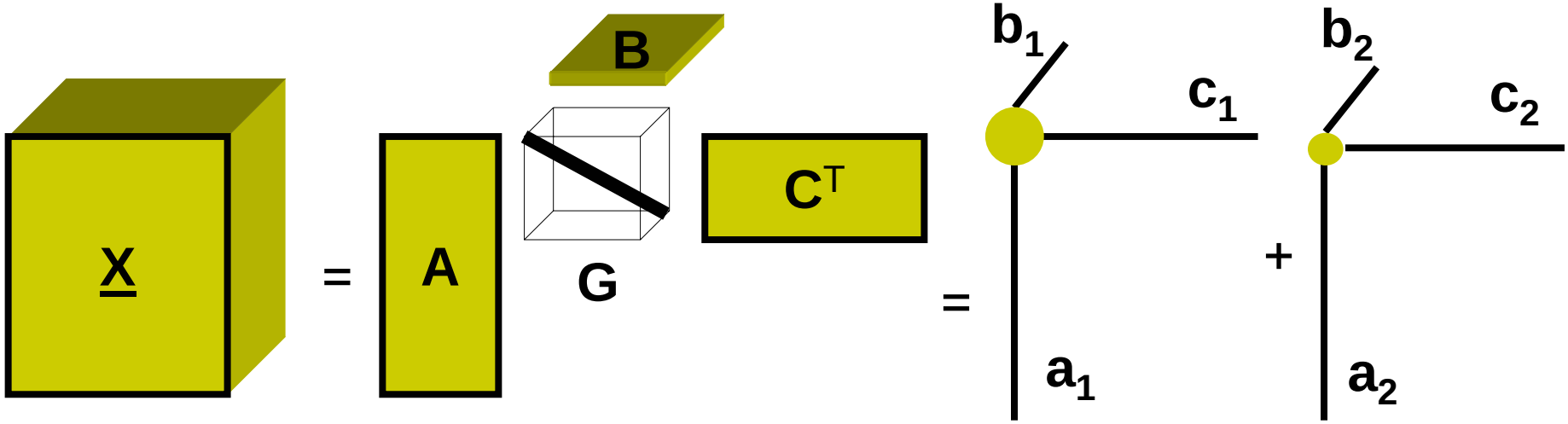


- MacGregor
  - Nonlinearities removed
  - Whole batch is considered a measurement
  - Off-line monitoring

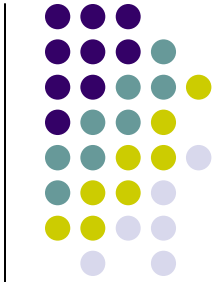


Extension of SVD to Parafac



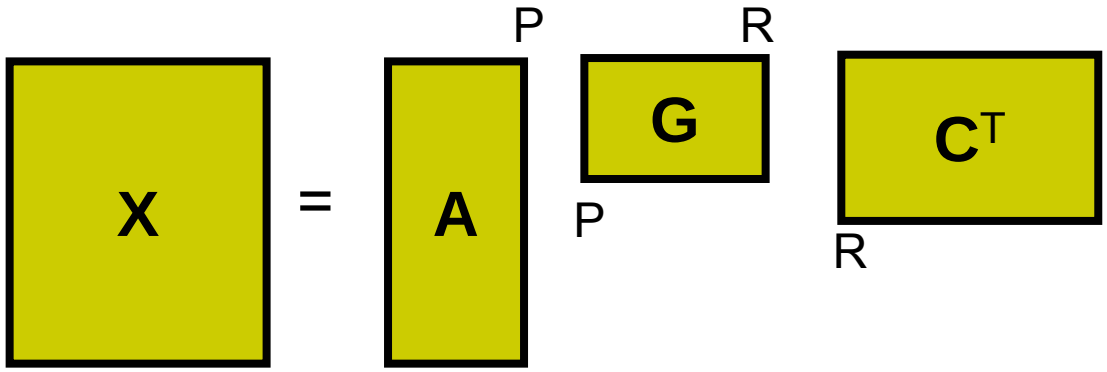


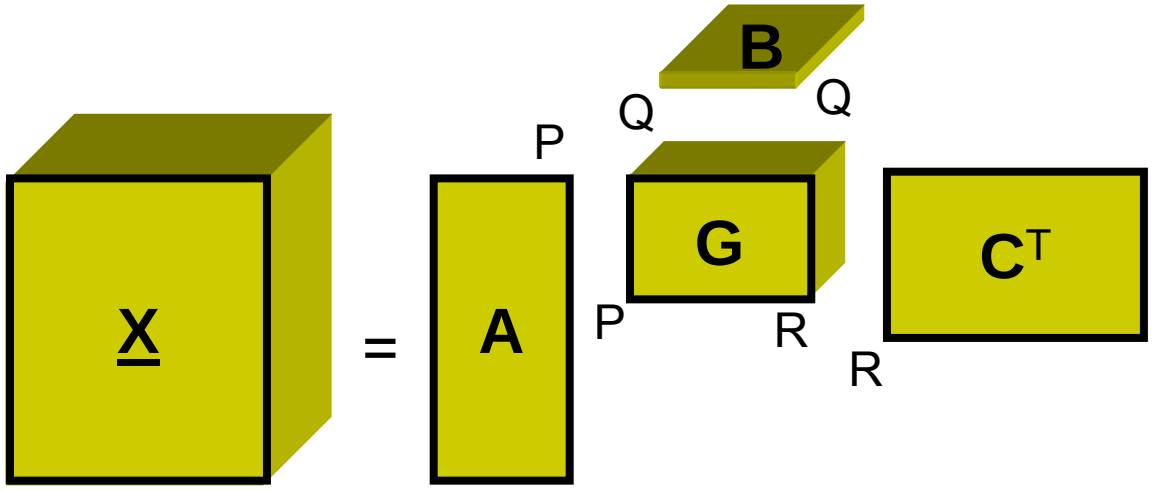
Parafac / Candecomp



- Parafac is not sequential
  - Need to re-estimate whole model when more components are calculated [no deflation].
- Parafac solution is unique
  - No rotational freedom
  - Changing parameters will reduce the fit.
  - **NB!** A PCA model is not unique
  - $X = T*P^T + E = T*R*R^{-1}*P^T + E = C*S^T + E$
  - Unique ≠ true

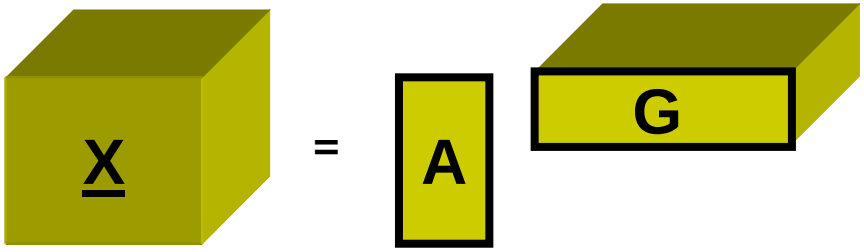
Extension of Two Mode component Analysis (TMCA)



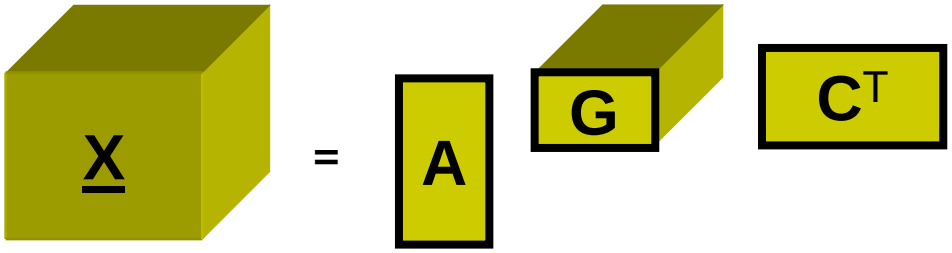


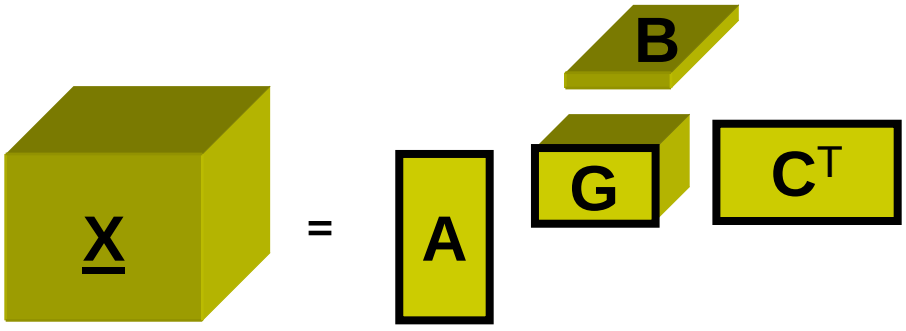
Tucker III

Tucker models

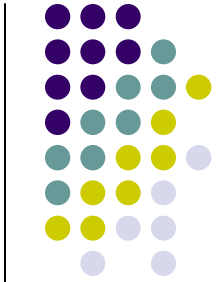
- Tucker I, 

Equals MPCA

- Tucker II, 

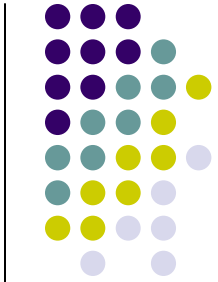
- Tucker III 

Tucker models



- Core array can be fully filled
- PxQxR triads (1,1,1 / 1,1,2 / 1,2,1 etc)
- Not unique rotational freedom
  - Components can be rotated towards orthogonality.
- Not sequential
- Restricted Tucker models can be developed when using prior chemical knowledge

Number of parameters

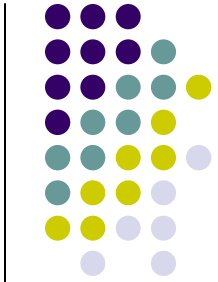


- X(IxJxK) example I=50, J=9, K=100,
  - P = Q = R = 3

- Parafac: Rx(I + J + K) 477
- Tucker3: PxI + QxJ + RxK + PxQxR 504
- MPCA: Rx(I + JK) 2850

- Fit MPCA > Parafac (Overfit?)

Soft models vs hard models



- Two-way bilinear model:
  - Beer's law

$$A_{i,\lambda} = \varepsilon_{1,\lambda} c_{i,1} + \varepsilon_{2,\lambda} c_{i,2} + e_{i,\lambda}$$

No orthogonal constraints

  - PCA

$$x_{ij} = t_{i1} p_{j1} + t_{i2} p_{j2} + e_{ij}$$

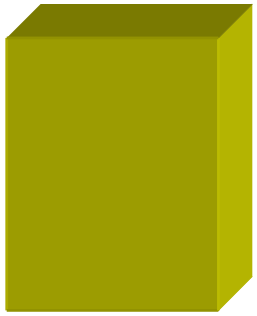
Orthogonal constraints

- Trilinear model:
  - Parafac
  - Fluorescence

$$x_{ijk} = a_{i1} b_{j1} c_{k1} + a_{i2} b_{j2} c_{k2} + e_{ijk}$$

No orthogonal constraints

Multiway Regression I



- Two step approach:

$$\tilde{X} = A\tilde{P} + E$$

$$\min_{A,\tilde{P}} \left\{ \left\| \tilde{X} - A\tilde{P} \right\|^2 \right\}$$

$$y = Ab + f$$

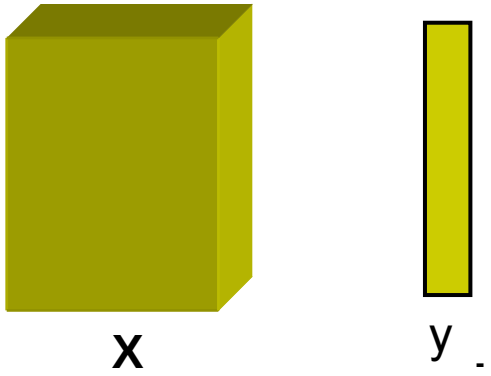
$$\min_{b} \left\{ \left\| y - Ab \right\|^2 \right\}$$

Decomposition of X to A and model $\tilde{P}$
Regression of y on A

$\tilde{P}$ Can be Parafac, Tucker, MPCA etc

No information of Y is used in the decomposition
Similar to PCR method

Multiway Regression II



- Direct approach

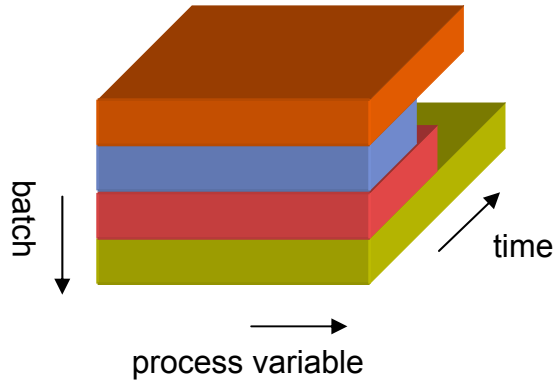
$$\tilde{X} = A\tilde{P} + E$$

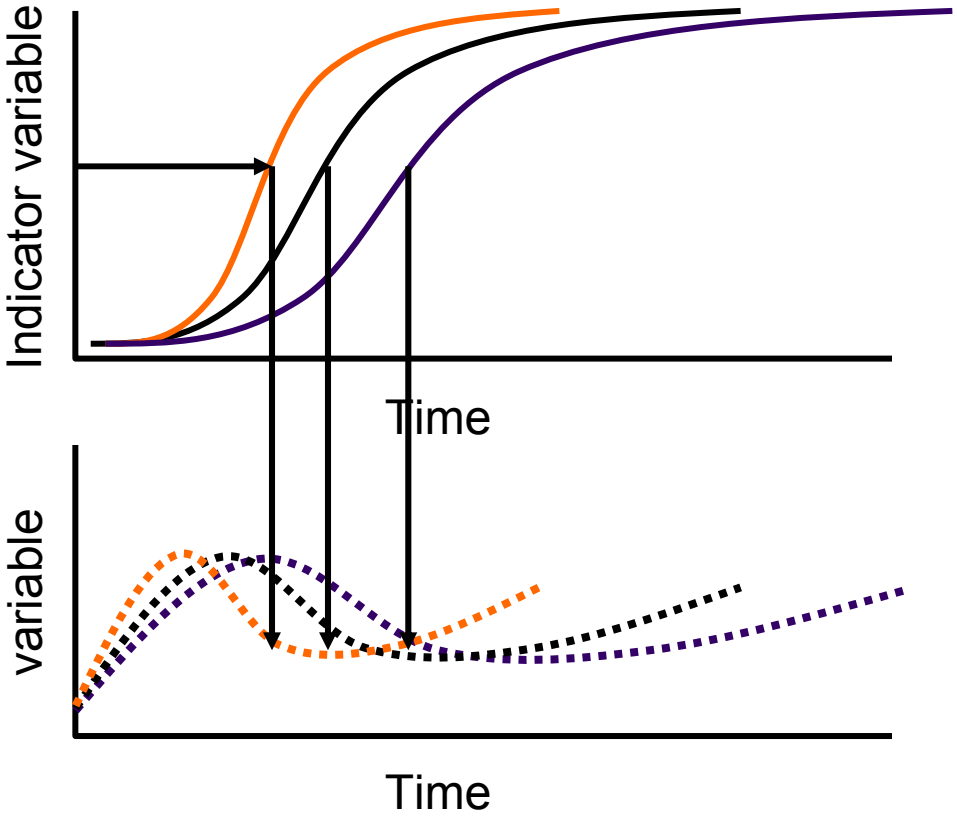
$$y = Ab + f$$

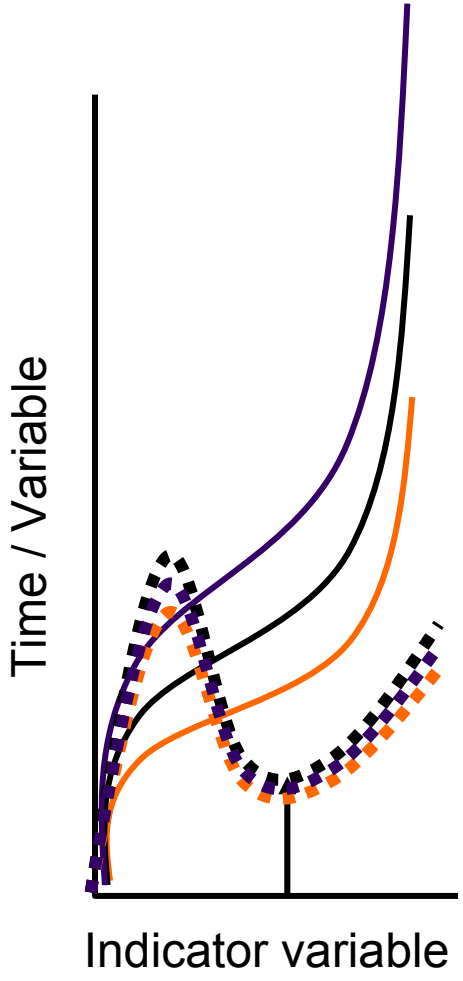
$$\min_{A,\tilde{P},b}\left\{\left\|\tilde{X} - A\tilde{P}\right\|^2 + \lambda\left\|y - Ab\right\|^2\right\}$$

Now X is decomposed with y in mind.
This leads to a not optimal decomposition of X but an improved fit of y.

When data are not exactly 3-way

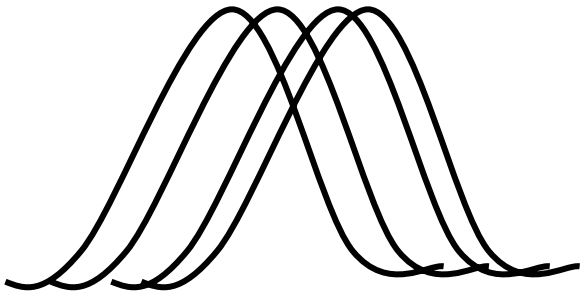




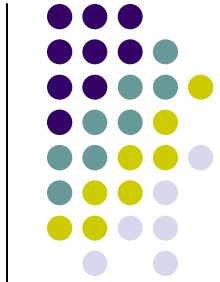


Alignment problems

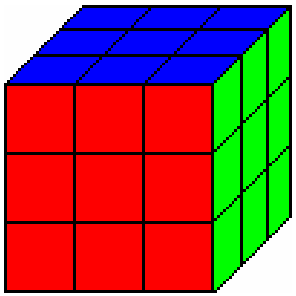
- Peakshifts in LCMS/GCMS



- Warping methods to align the peaks
  - Dynamic Time Warping
  - Correlation optimized warping



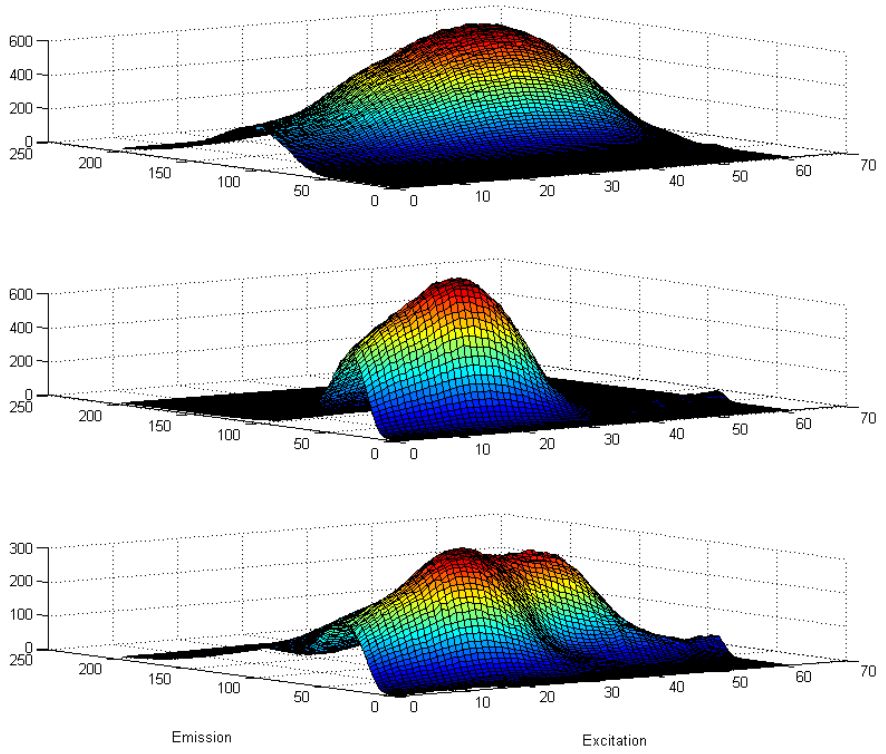
Three-way Applications



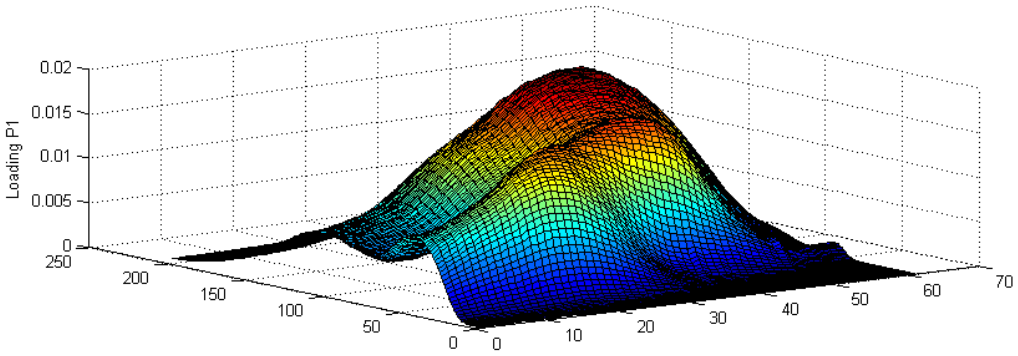
Fluorescence data

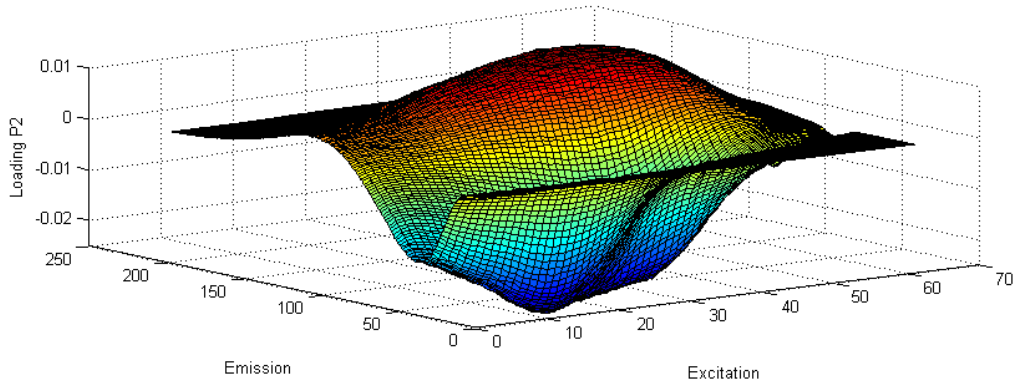
- 5 samples with varying concentration of tyrosine, tryptophan and phenylalanine dissolved in phosphate buffered water.

- Excitation wavelength: 240 – 300 nm
- Emission wavelength: 250 – 450 nm



Unfold PCA model of Fluorescence data



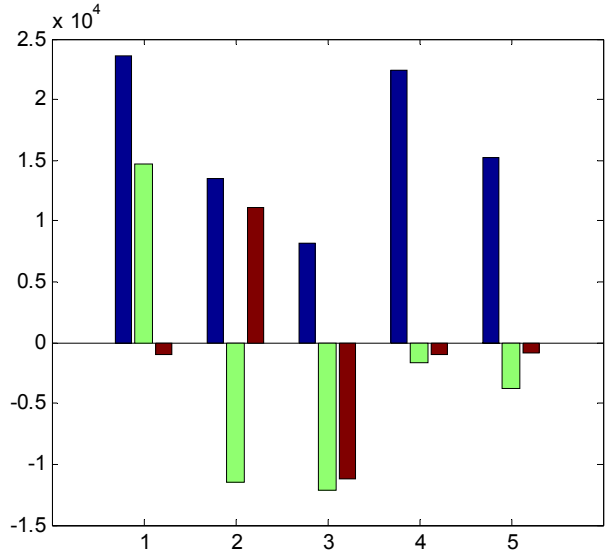


99.97% explained with 3 PC's

Loadings refolded into Excitation / Emission form

Overfit of data:

Loading 2 has negative parts. This is not according fluorescence theory.



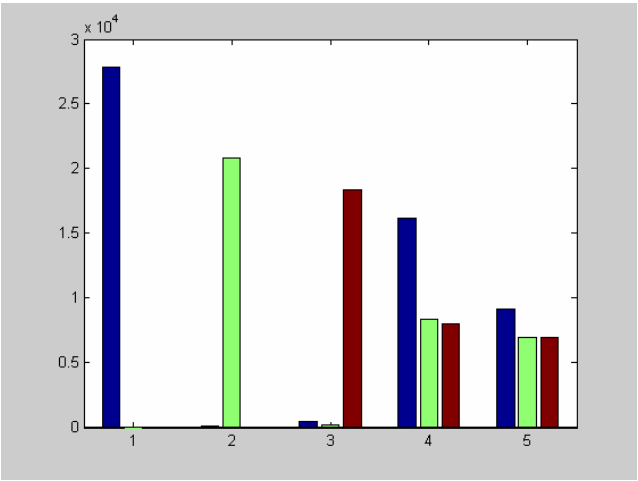
Parafac model of Fluorescence data

99.93% explained variation: Good Fit

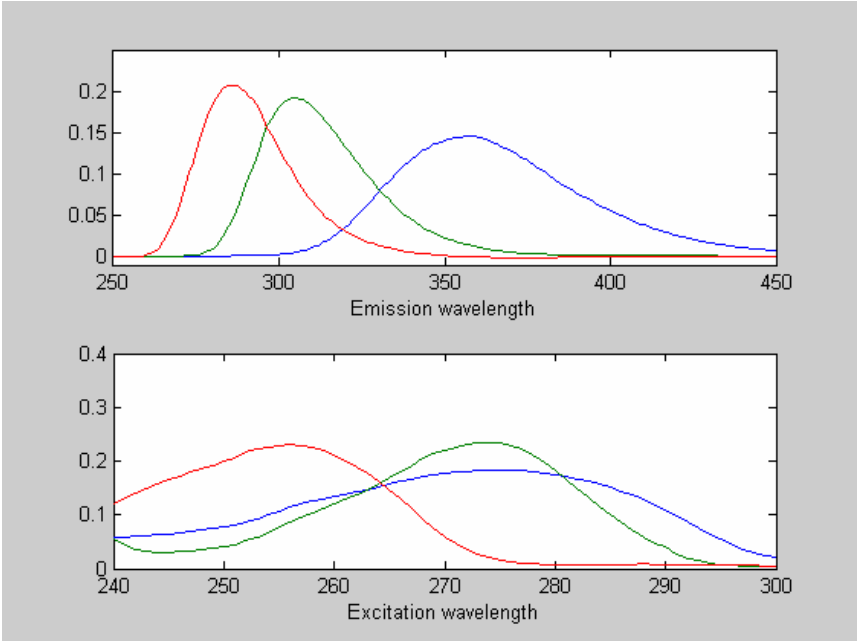
Loadings are very well interpretable.

Intensity in A mode can be related to concentration

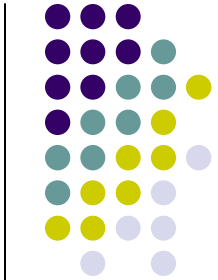
A mode



B and C mode



Fluorescence data



$$I_{\lambda Em,\lambda Ex,k} = a_{\lambda Em1} b_{\lambda Ex1} c_{k1} + a_{\lambda Em2} b_{\lambda Ex2} c_{k2} + a_{\lambda Em3} b_{\lambda Ex3} c_{k3} + e_{ijk}$$

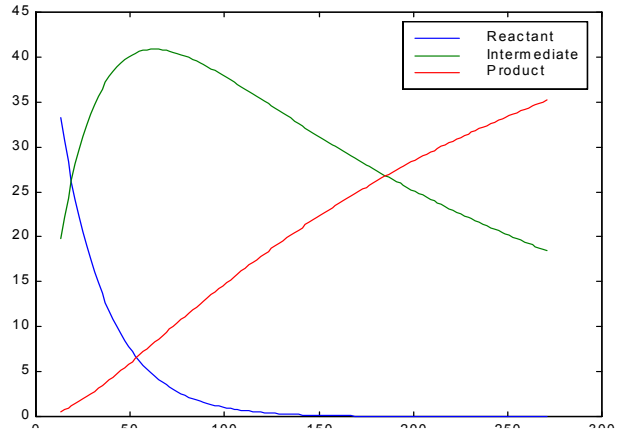
Florescence data perfectly fits the trilinear model that is applied by Parafac

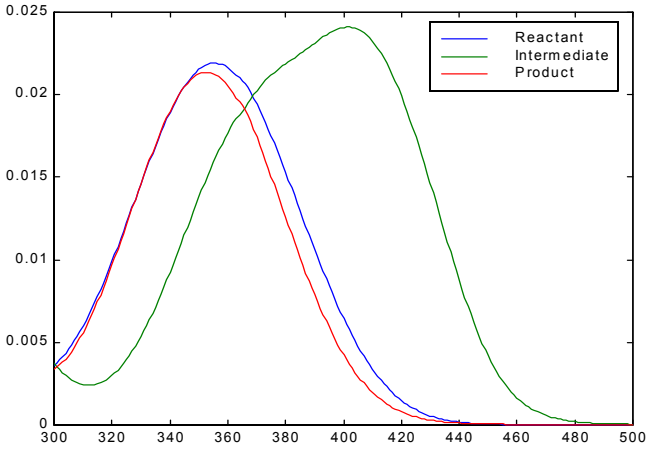
Due to uniqueness property of Parafac, the loadings found will perfectly resemble the Emission spectra and Excitation spectra of the three compounds in de mixtures.

This is a nice example of Mathematical chromatography

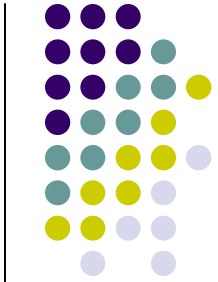
Batch reaction monitoring

- **Pseudo-first-order reaction:**
  A + B → C → D + E
- **UV-Vis spectrum (300-500nm) measured every 10 seconds.**
- **Obeys Lambert-Beer law**
- **35 NOC batches. $\underline{X}$ ($35 \times 201 \times 271$)**
- **In addition, some disturbed batches were measured**
  - **pH disturbance during the reaction**
  - **Temperature change**
  - **Impurity**



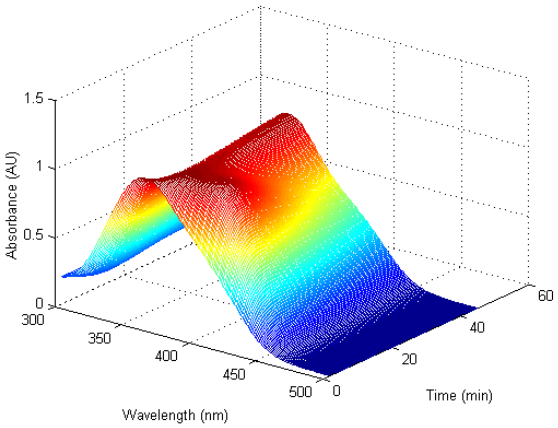


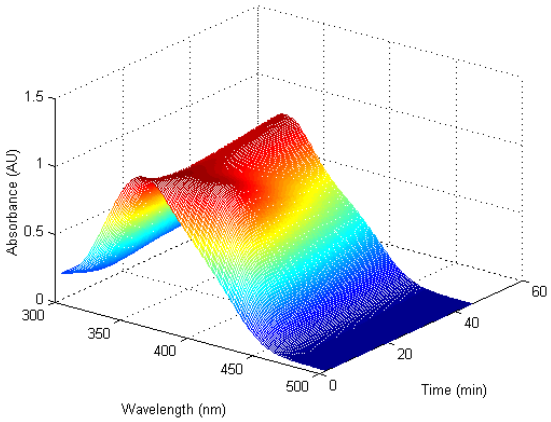
Aims and goals of research I

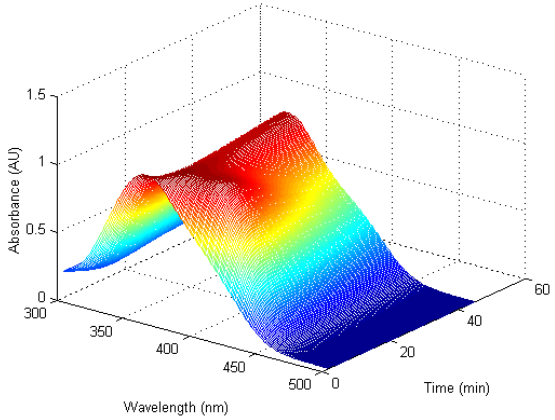


- **Data modelling**:
  - Improve understanding of process by interpretation of model parameters
- **Analysis of historical batches**:
  - Are the current process measurements able to distinguish between ‘good’ and ‘bad’ batches?
- **On-line monitoring**:
  - Rapid fault detection
  - Easier fault diagnosis: what is the cause of the fault?
  - Prediction of batch duration

Aims and goals of research II



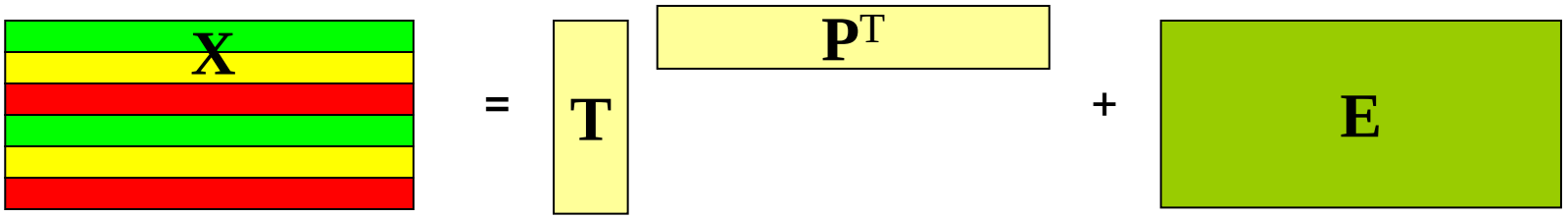




Which batch is different ?

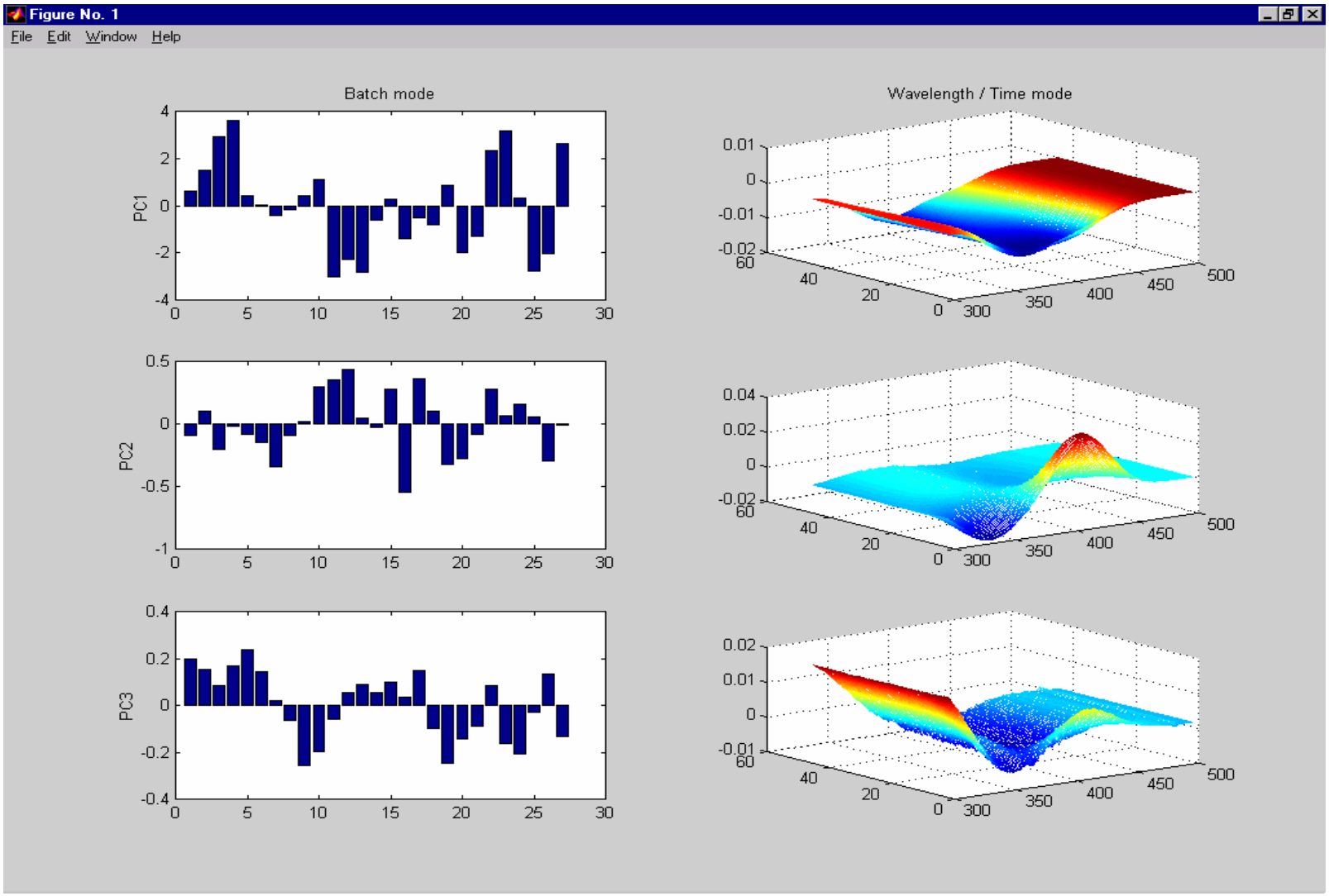
Unfold PCA model

- Unfold keeping the batch direction (IxJK)



$$\mathbf{x}_{i,jk} = \sum_{r} \mathbf{t}_{i,r}\mathbf{p}_{jk,r} + \mathbf{e}_{i,jk}$$

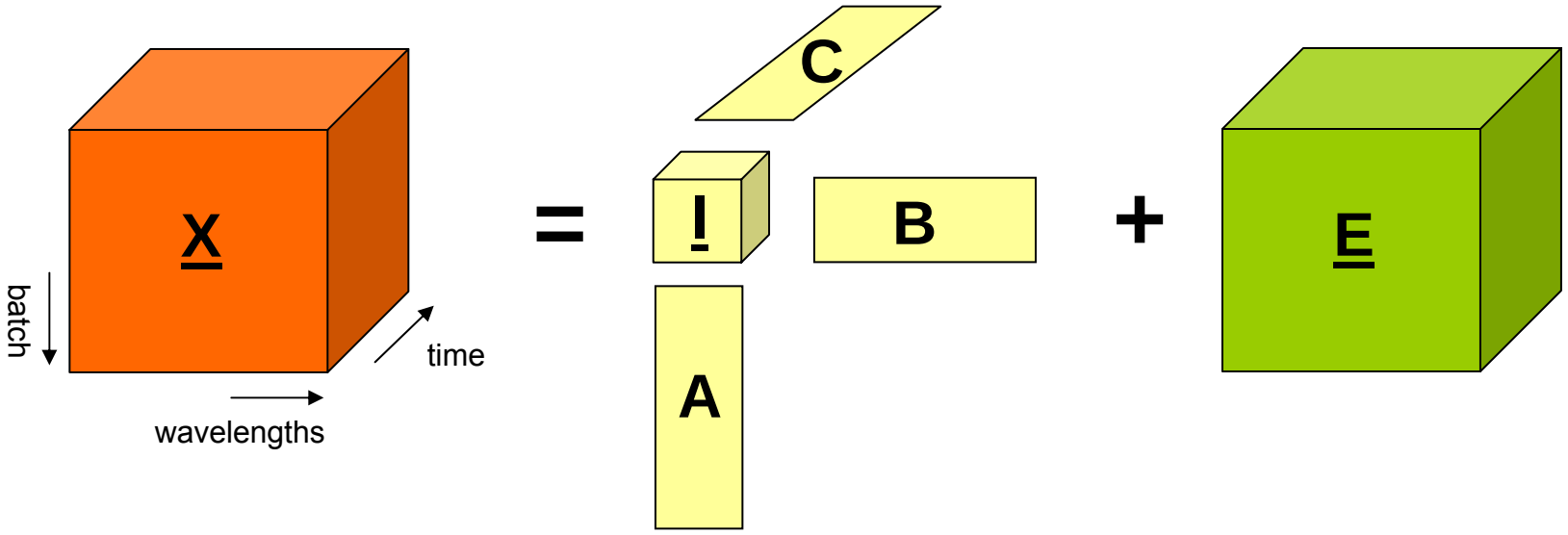
Unfold PCA model



Many parameters estimated, likely to overfit the data

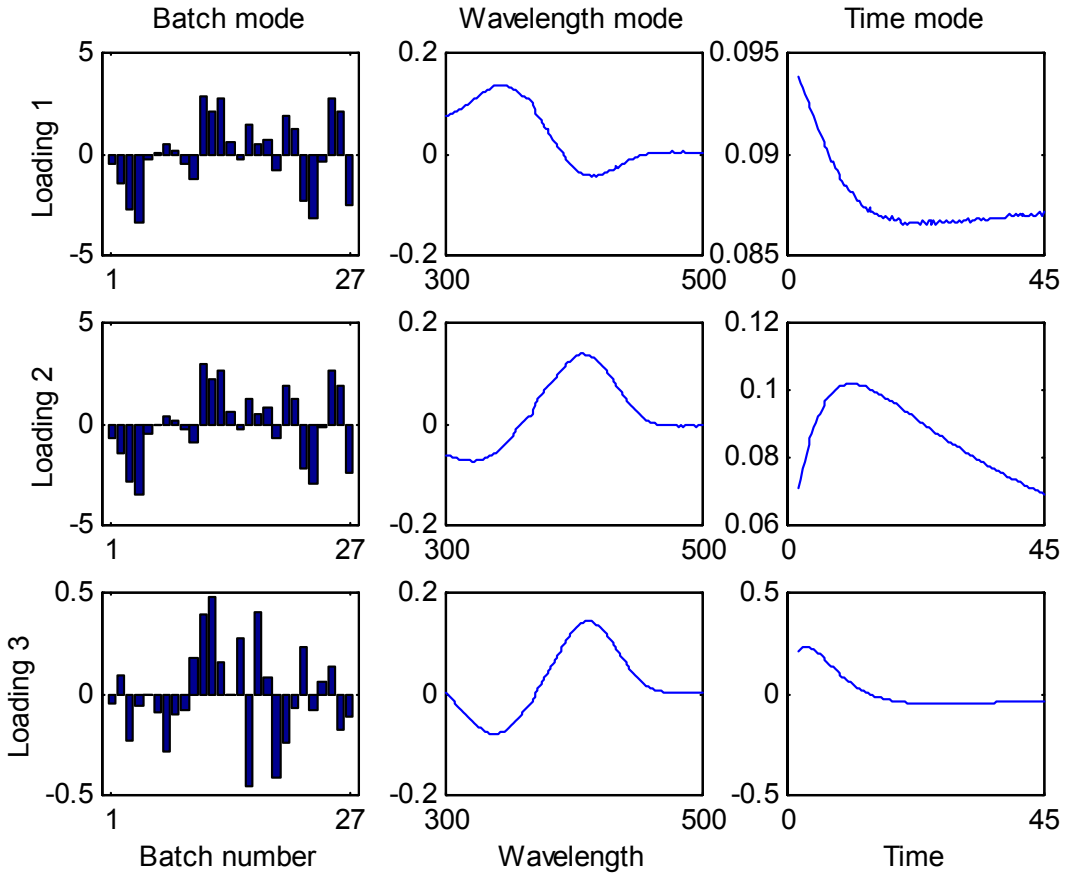
Unrestricted Parafac model

- The simplest three-way model is the PARAFAC model:



Unrestricted Parafac model

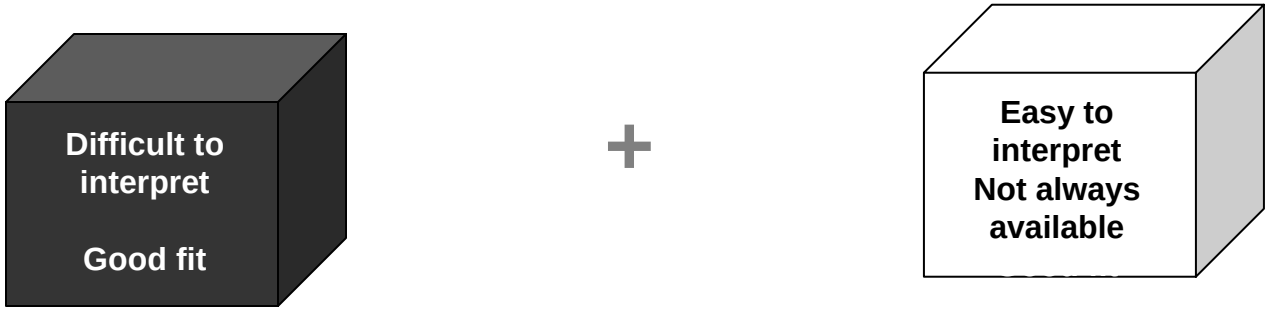
- Loadings are highly correlated - solution may be unstable.
- Model is difficult to interpret.
- 99.4% fit
- *Can external knowledge of the process be used to improve the model?*



Grey Modelling of batch data

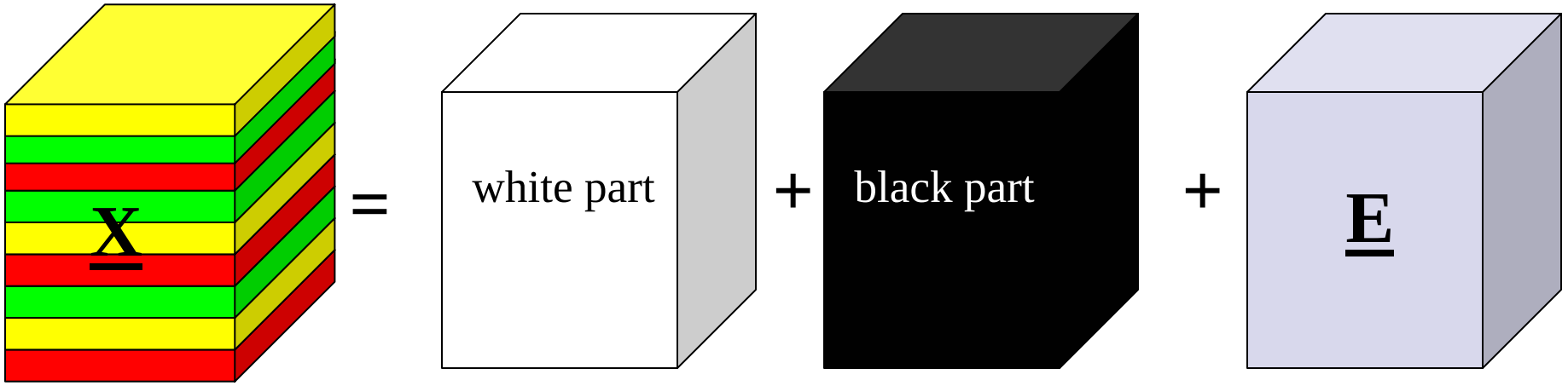
‘Black-box’ or ‘soft’ models are empirical models which aim to fit the data as well as possible *e.g.* PCA, neural networks.

‘White’ or ‘hard’ models use known external knowledge of the process *e.g.* physicochemical model, mass-energy balances.



‘Grey’ or ‘hybrid’ models combine the two.

Modelling batch data



Total variation

Systematic variation due to known causes

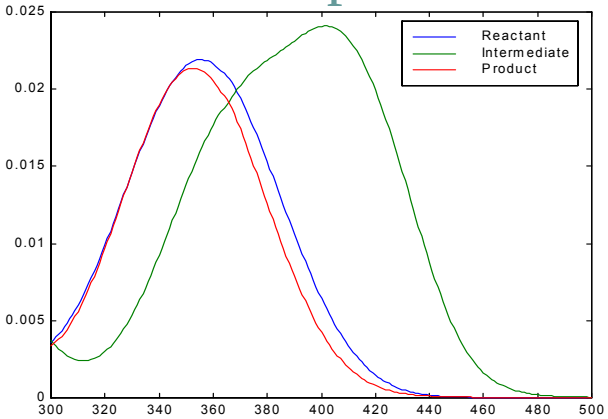
Systematic variation due to unknown causes

Unsystematic variation

External information

- Incorporating external information can
  - increase model interpretability
  - increase model stability

Pure Spectra



Reaction kinetics

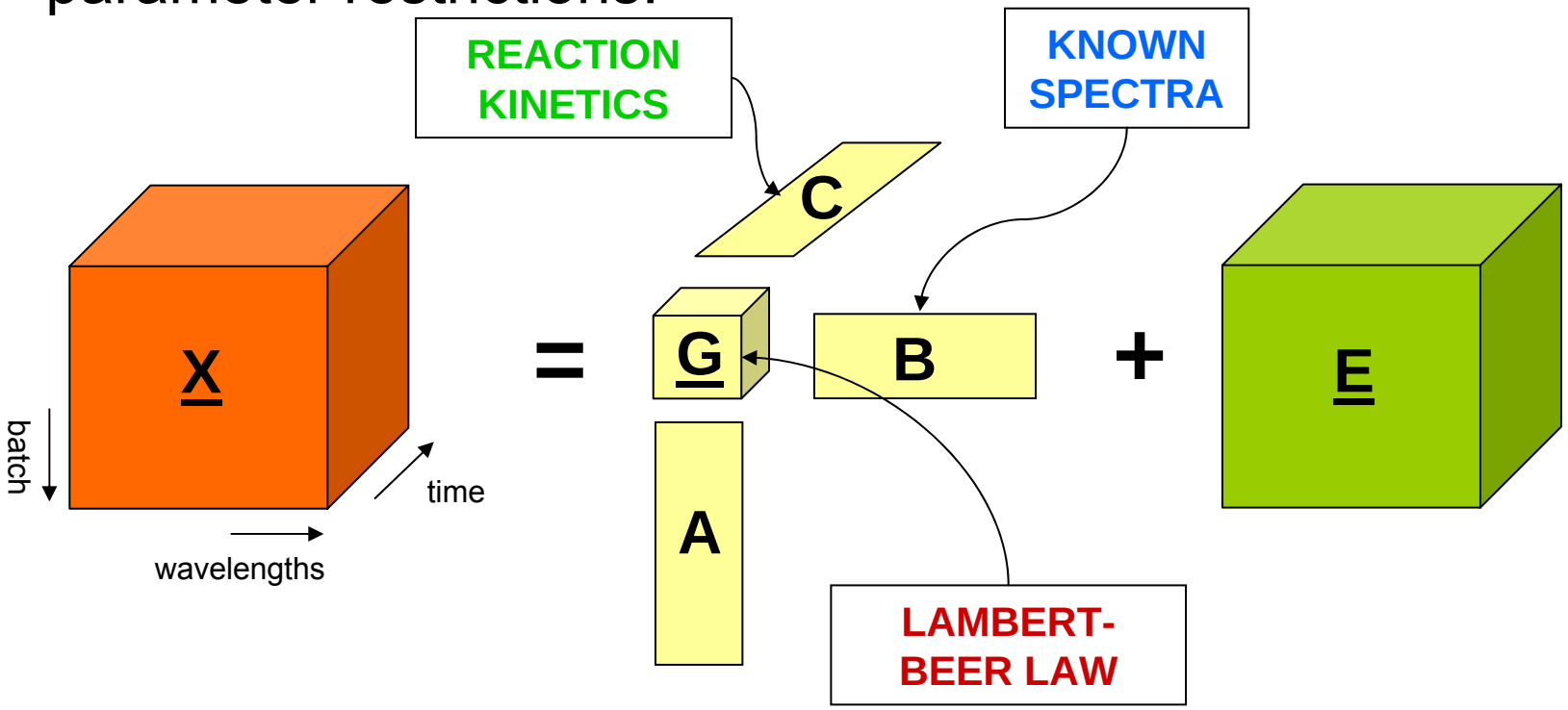
$$[\mathrm{A}]_t = [\mathrm{A}]_0 e^{-k_1 t}$$

$$[\mathrm{C}]_t = \frac{k_1[\mathrm{A}]_0}{k_2 - k_1}\left(e^{-k_1 t} - e^{-k_2 t}\right)$$

$$[\mathrm{D}]_t = [\mathrm{A}]_0 - [\mathrm{A}]_t - [\mathrm{C}]_t$$

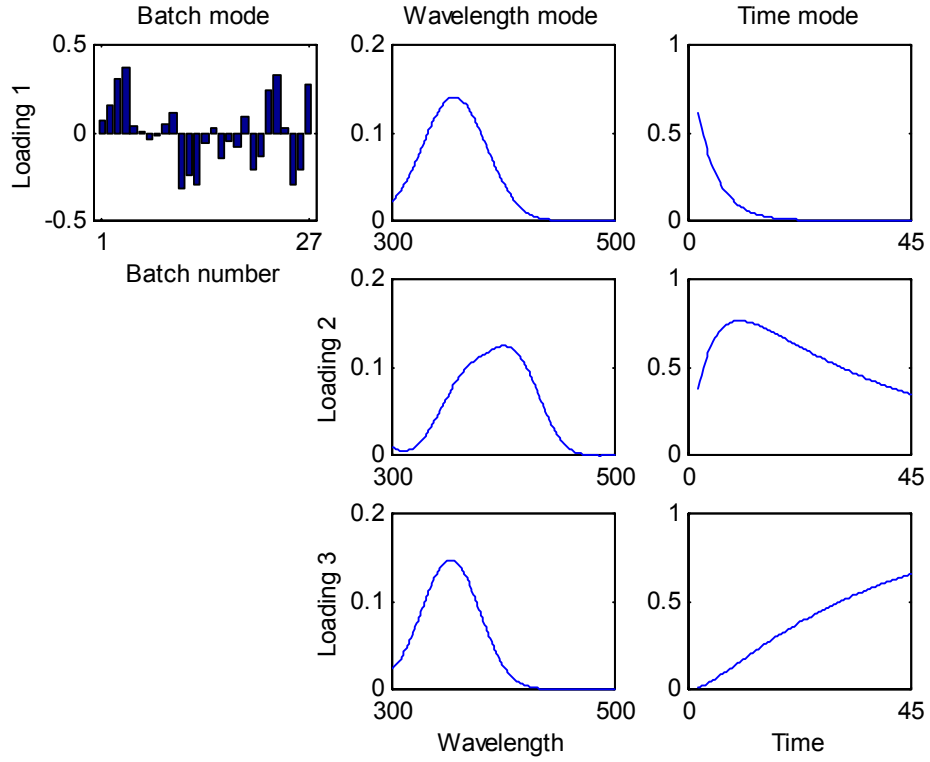
Restricted ‘white’ model

- External information is introduced in the form of parameter restrictions:

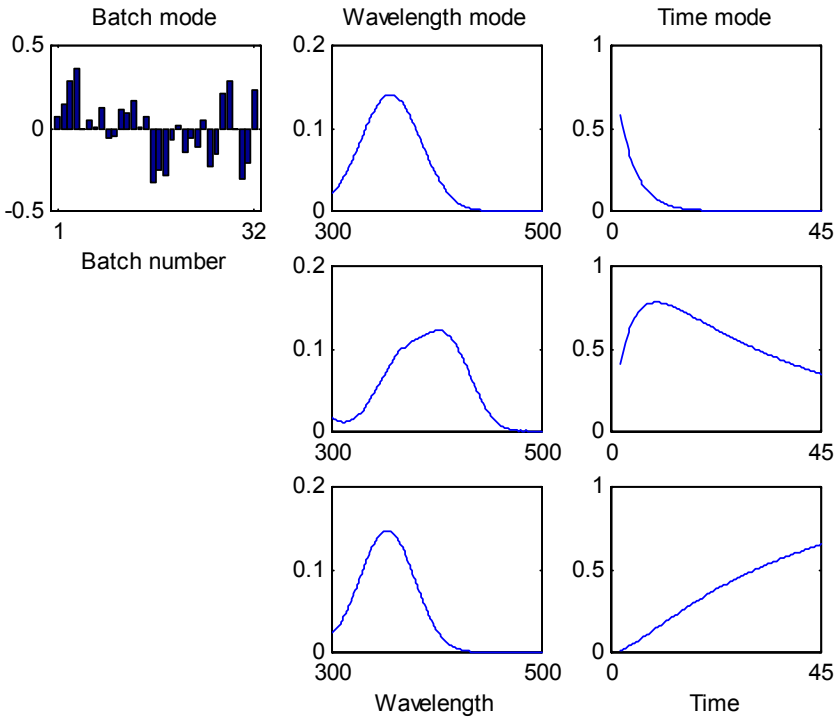


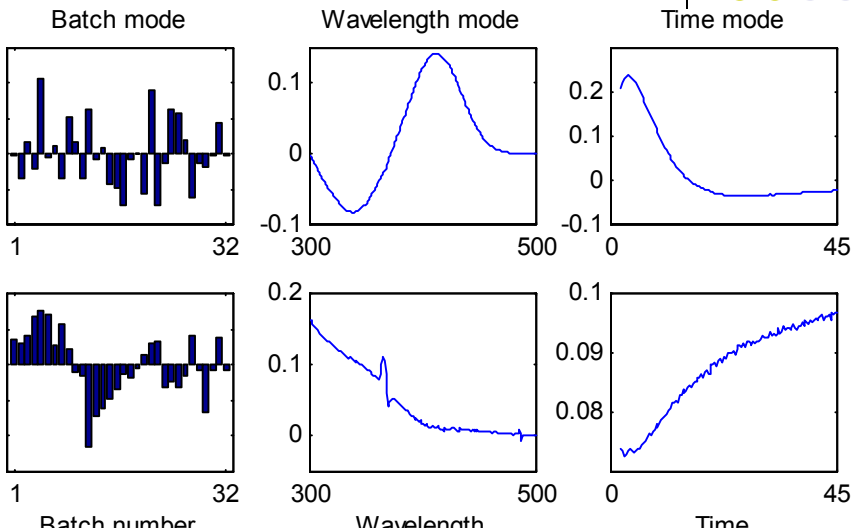
Restricted Tucker model

- Model is stable.
- 97.6% fit - lower than for black model
- *Some systematic variation in the data is left unexplained by this model.*



Grey model



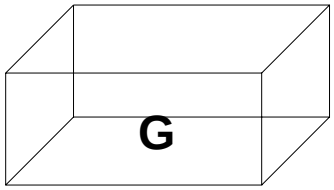


White components describe known effects

Black components can be interpreted

- 99.8% fit (corresponds well with estimated level of spectral noise of ≈ 0.13%)

Core array of restricted Tucker model

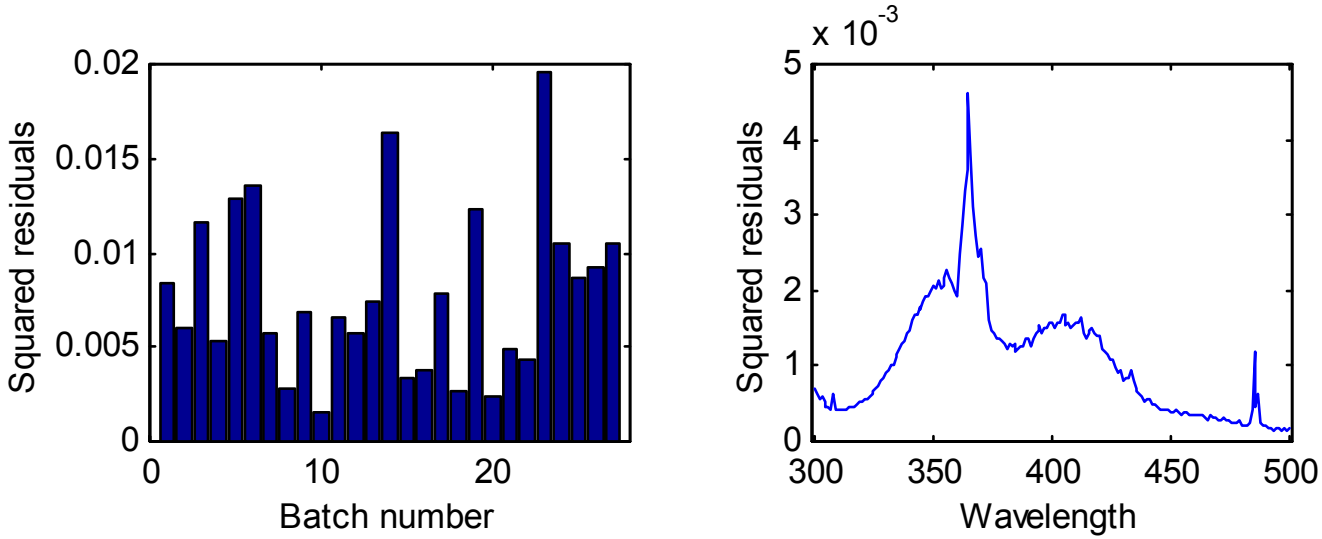


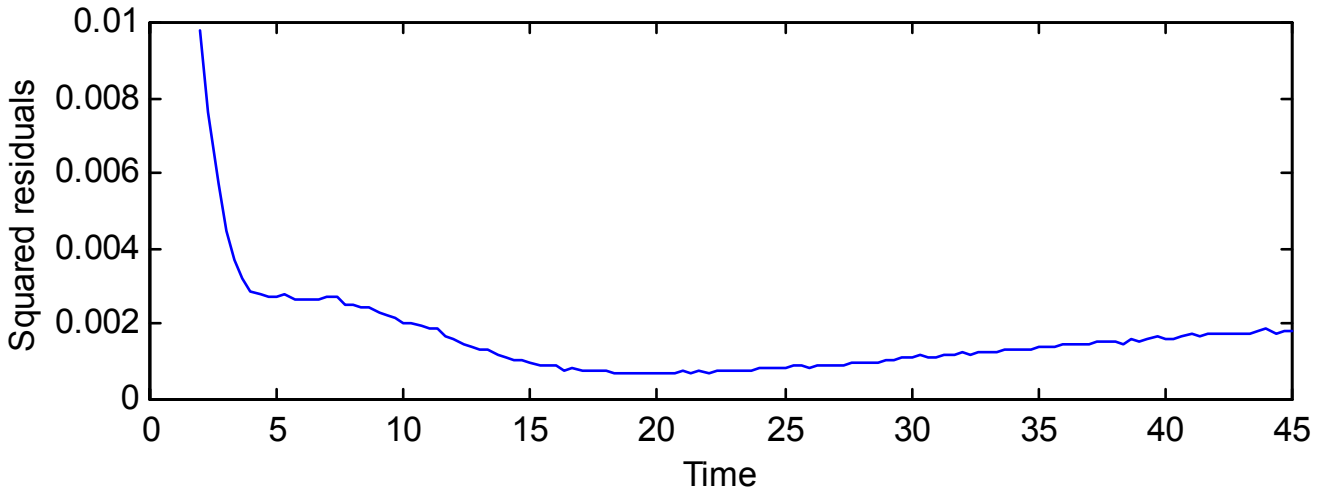
3x5x5
core array

g_{111}	0	0	0	0
0	g_{122}	0	0	0
0	0	g_{133}	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	g_{244}	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	g_{355}

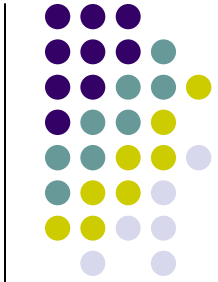
- Only combinations:
  - g_{111}, a_1, b_1, c_1
  - g_{122}, a_1, b_2, c_2
  - g_{133}, a_1, b_3, c_3
  - g_{244}, a_2, b_4, c_4
  - g_{355}, a_3, b_5, c_5

Grey model residuals





Properties of grey models



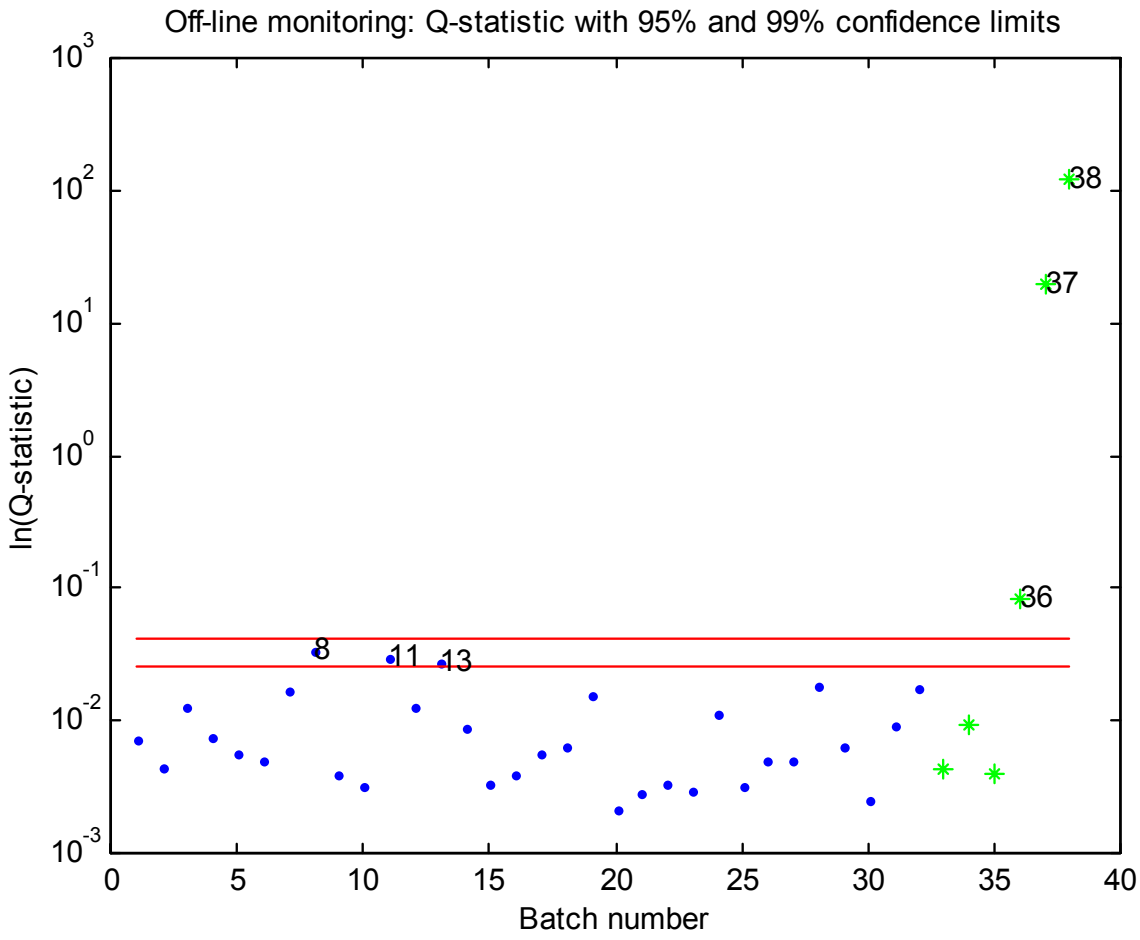
- White and black model parts can be calculated
  - simultaneously (*via* restricted core matrix) with better % fit
  - sequentially with better diagnostics - allows partitioning of variance

$$\|\mathbf{X}\|^2 = \|\mathbf{X_w}\|^2 + \|\mathbf{X_b}\|^2 + \|\mathbf{E}\|^2$$

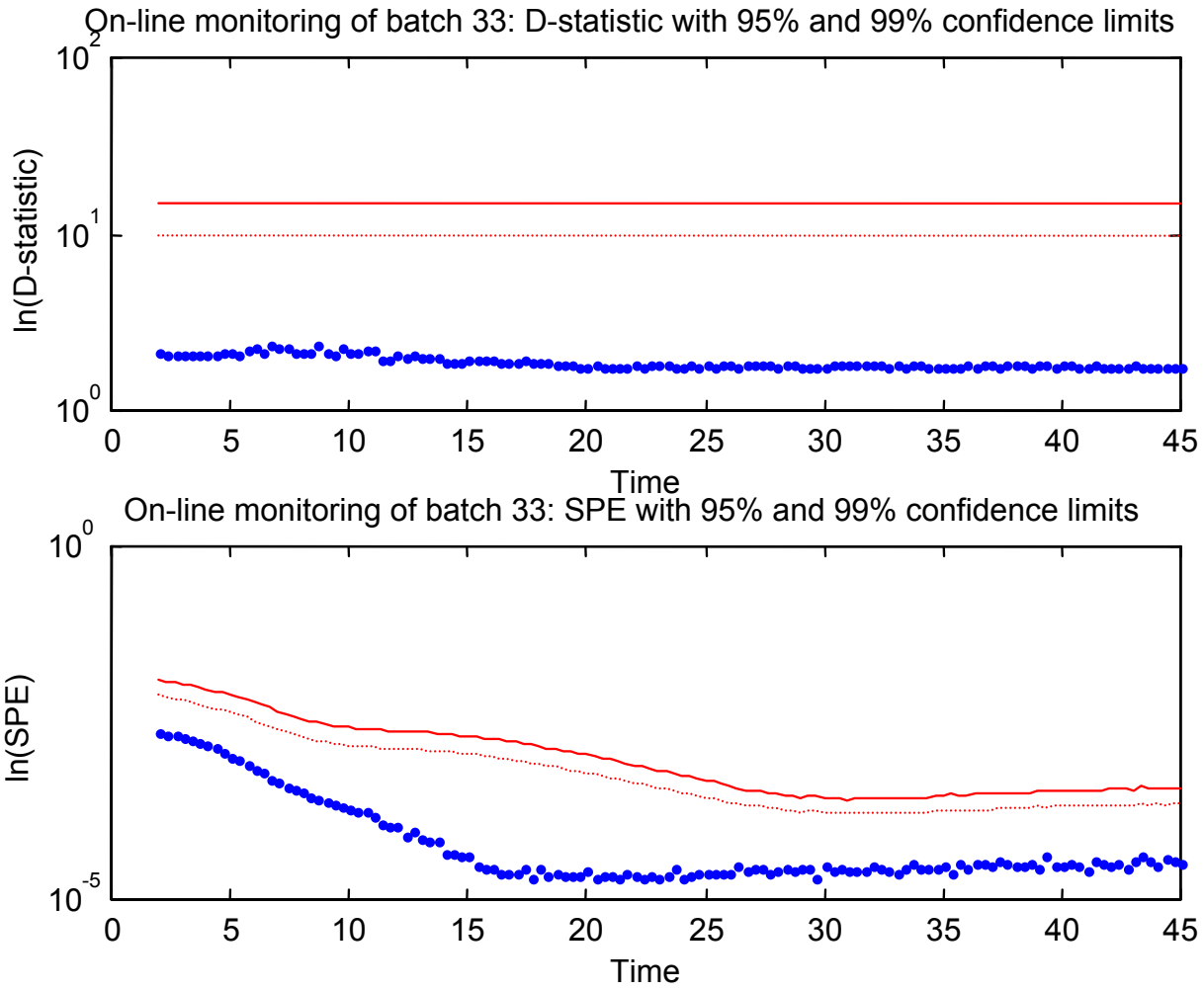
    - 100% = 97.1% + 1.9% + 0.2%
  - simultaneously but with orthogonality restrictions which also allow partitioning of variance

Off-line batch monitoring

- NOC: # 1:32
- Validation: # 33-35
- pH Disturbed: # 36
- Temp. problem # 37
- Impurity # 38

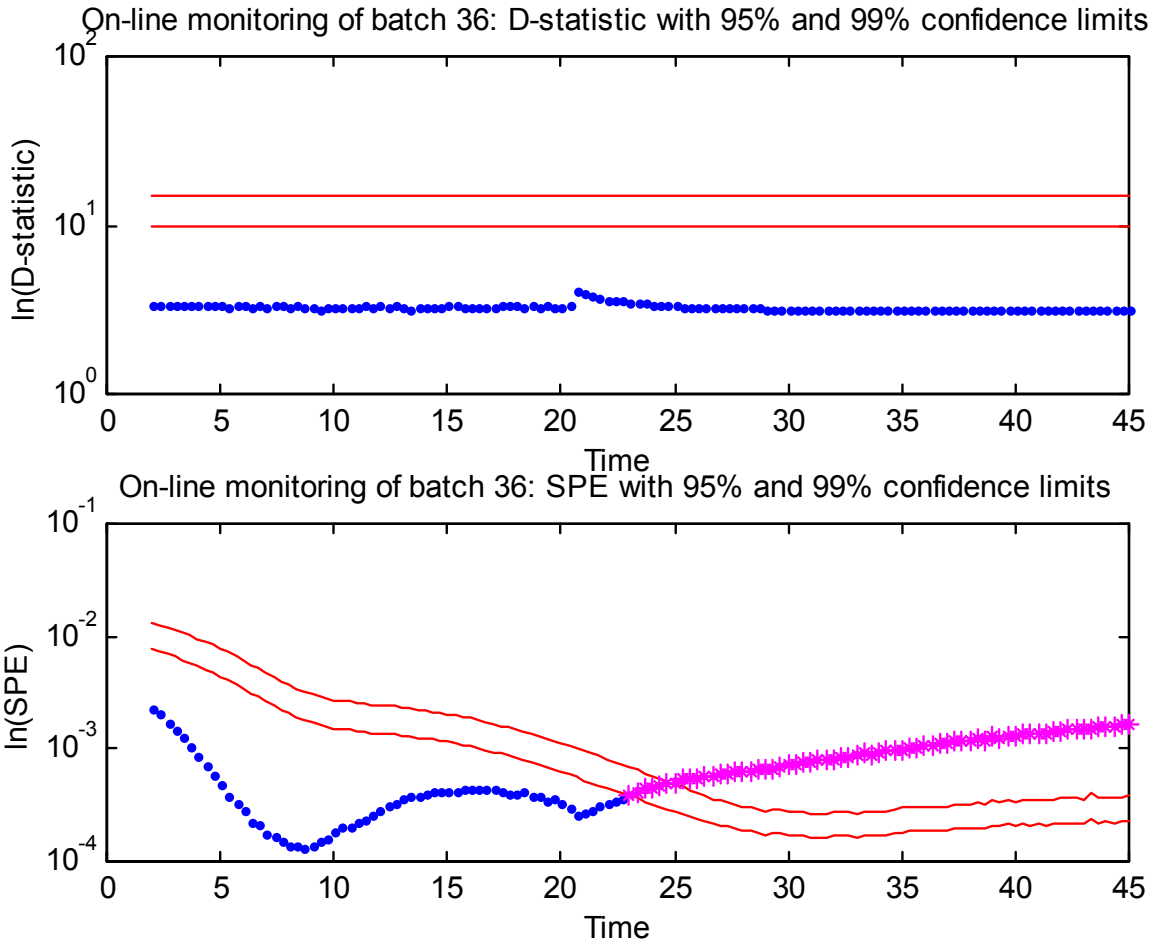


On-line monitoring of a validation batch



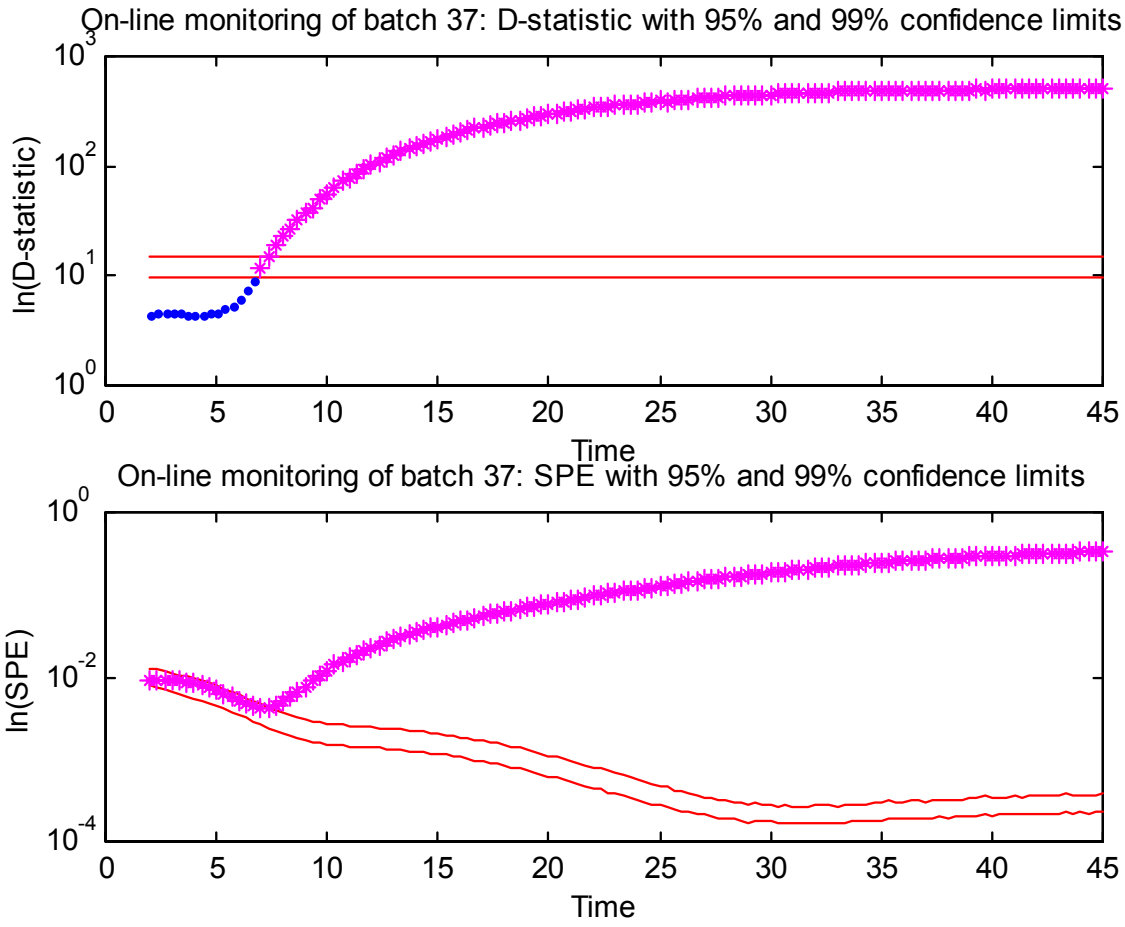
On-line monitoring of the pH disturbed batch

- After 23 minutes SPE goes outside control limits
- pH was disturbed after 21 minutes
- Only small change in D-statistic



On-line monitoring of the temperature disturbed batch

- Temperature slowly decreasing from start of reaction
- Rate constant k_1 lower than usual.
- Contribution plot shows difference spectrum between reactant (too high) and intermediate (too low)



Want to know more
Look at Rasmus Bro's website

KVL

Quality and Technology
Department of Food Science

Latest news: **New poster page**

- Home
- Staff
- Research
- Teaching
- About

Download:
- Algorithms
- Courses
- Data Sets
- References
- Theses

Search Site:

Odin
data • information • control

Build Your Food

CampusNet

IFV IntraNet

Rasmus Bro

ᚱᚨᛋᛗᚢᛋ : ᛒᚱᚢ

Professor
Chemometrics Group, Dept. of Dairy and Food Science
The Royal Veterinary and Agricultural University
Rolighedsvej 30, DK-1958 Frederiksberg C
Phone: (+45) 35 28 32 96, Fax: (+45) 35 28 32 45
E-mail: Rasmus Bro

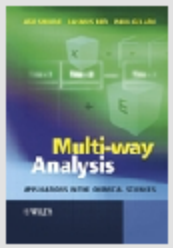
About Rasmus Bro

Curriculum vitae
Optima X
Presentations
Chemometrics World

Multi-way meeting
In July 2004 a workshop was held in Palo Alto bringing together researchers from math, stat, chemometrics, DSP, psychometrics and other areas. See more & more

Multi-way Analysis

Want to learn multi-way analysis?
Get a free monograph. Or pay!



The N-way Toolbox for Matlab
Triple-SPICE
Three-mode Company

Chemometrics

Literature database (incl. abstracts)
Links to chemometrics
MATLAB code
Data sets
The PLS_Toolbox