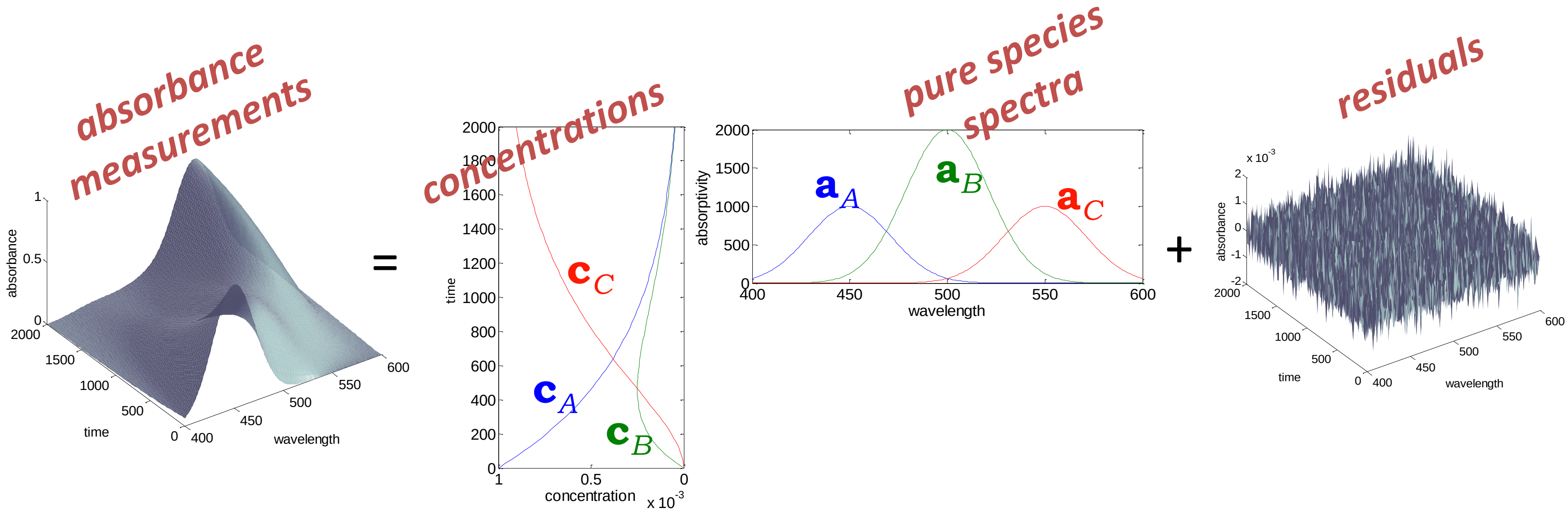
A “not exhaustive” overview of chemometric methods

a practical guide for data analysis in chemistry

Cap Sébastien

28 November 2014



What is Chemometrics?

Chemometrics is the science of extracting information from chemical systems by **data-driven means**. It is a highly interfacial discipline, using methods frequently employed in core data-analytic disciplines such as

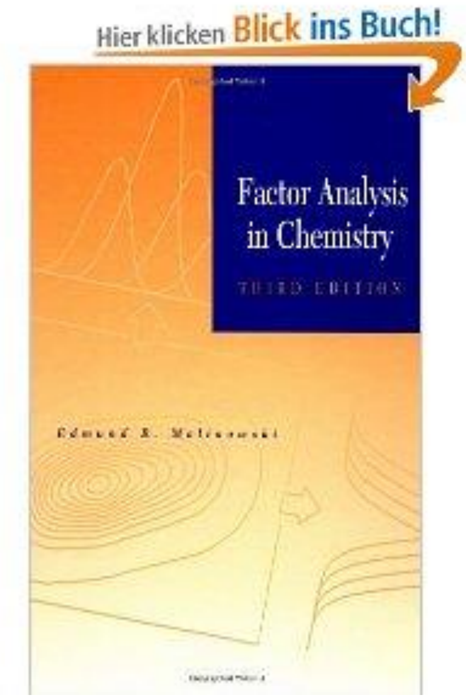
multivariate statistics, applied mathematics, and computer science,

in order to address problems in

chemistry, biochemistry, medicine, biology and chemical engineering.

In this way, it mirrors several other interfacial '-metrics' such as psychometrics and econometrics.

Reading Material



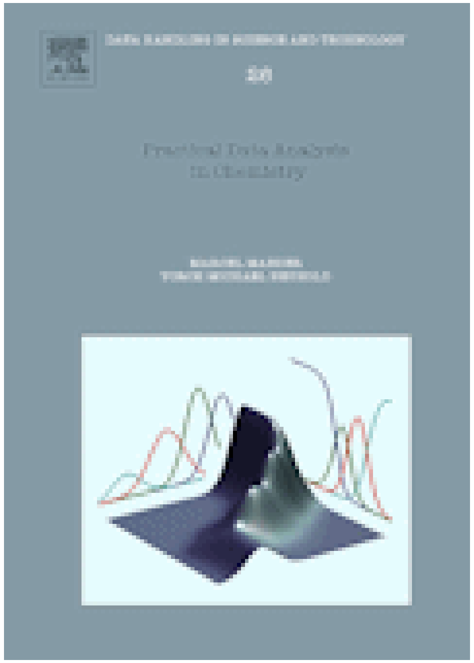
Factor Analysis in Chemistry, 3rd Edition

Edmund R. Malinowski

ISBN: 978-0-471-13479-4

432 pages

March 2002



Practical Data Analysis in Chemistry

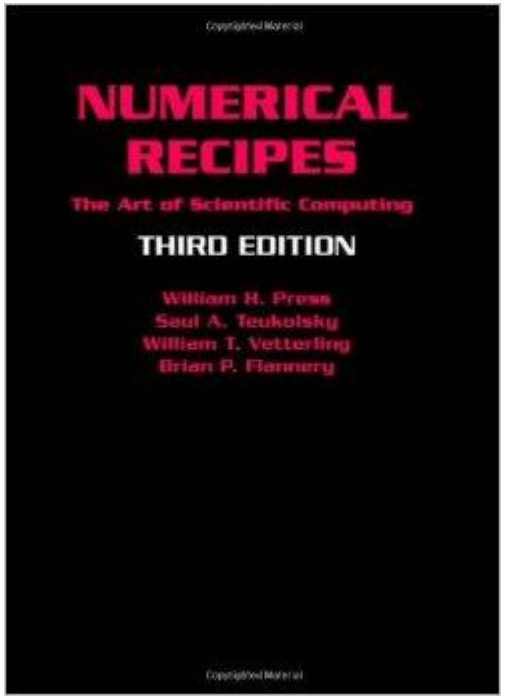
ByMarcel Maeder

Yorck-Michael Neuhold

Published: July 2007

Imprint: Elsevier

ISBN: 978-0-444-53054-7



Numerical Recipes 3rd Edition: The Art of Scientific Computing Hardcover – September 10, 2007

ISBN-13: 978-0521880688

1) Wikipedia

Table of content

1) Introduction / Framework

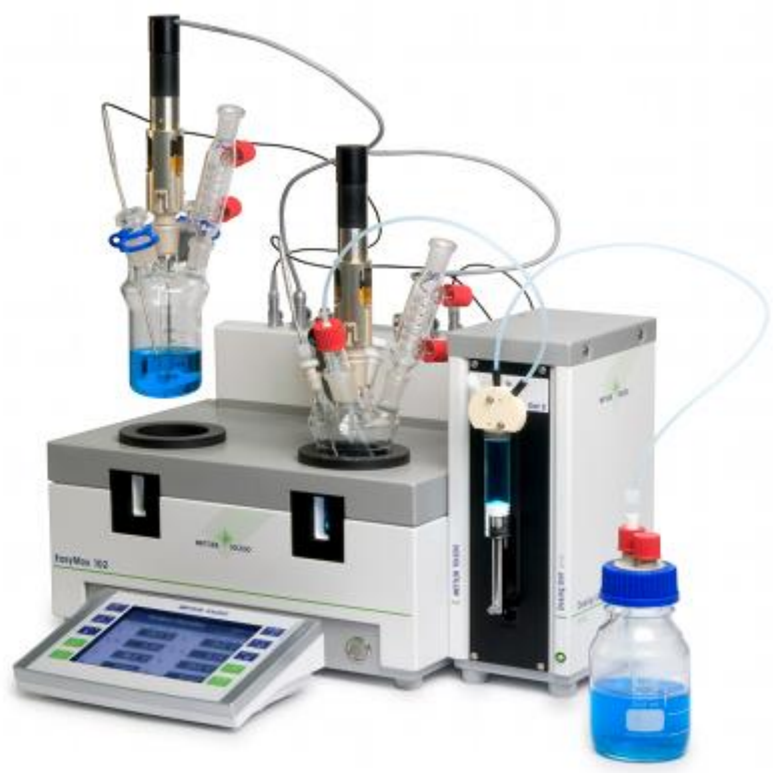
2) Instrumentation / Design of Experiment DoE

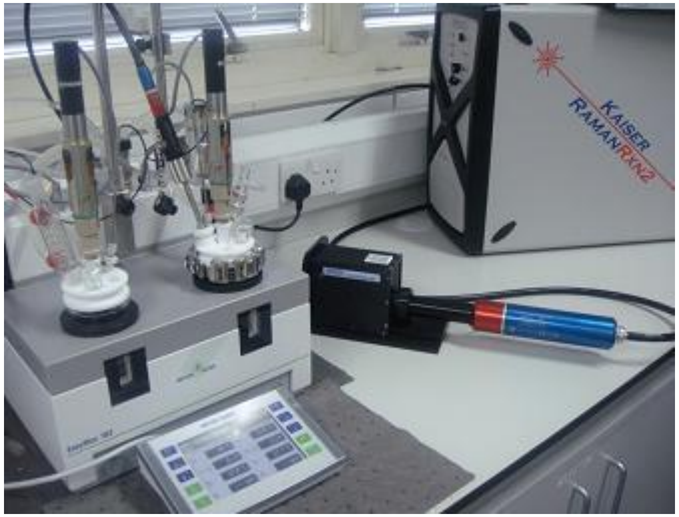
3) Kinetic soft modelling methods
(FA, EFA, MCR-ALS)

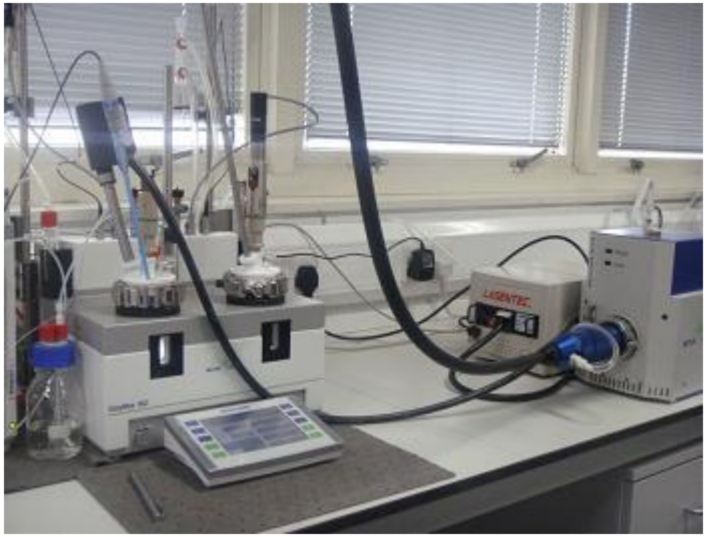
4) Kinetic hard modelling
Kinetic fitting of spectroscopic data with non-linear regression

Framework

Reactor / Instrumentation



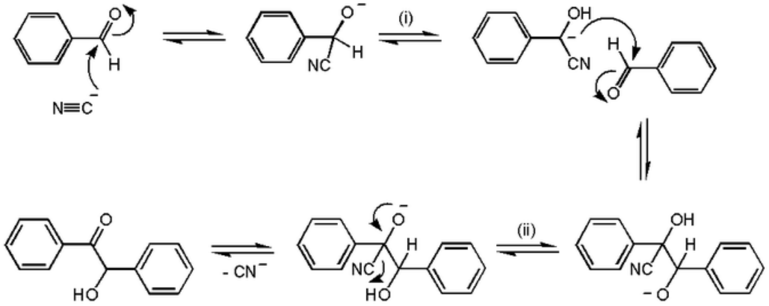




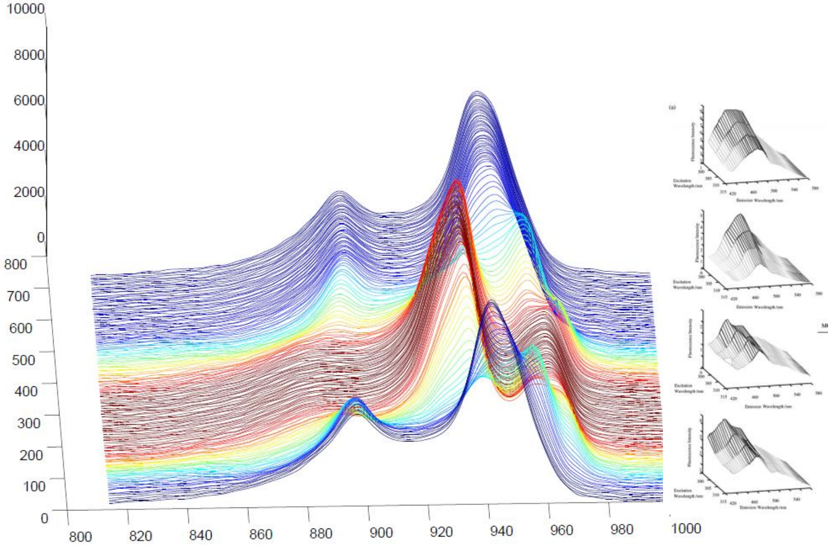
Framework

Chemistry

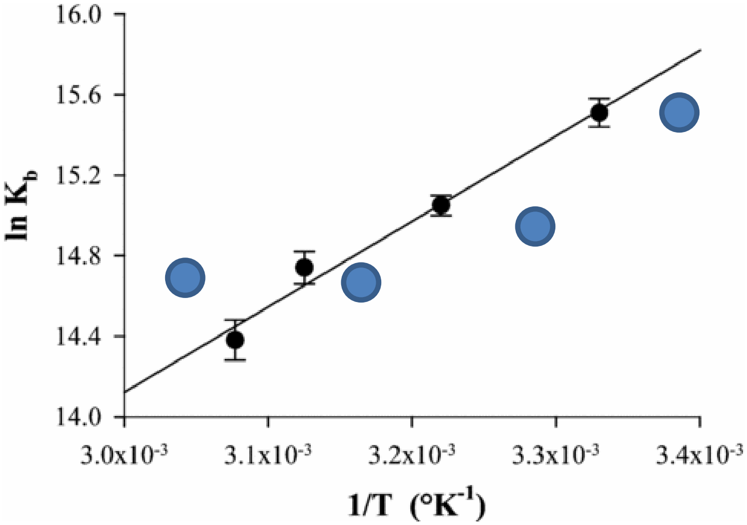
Plausible mechanism?



Data



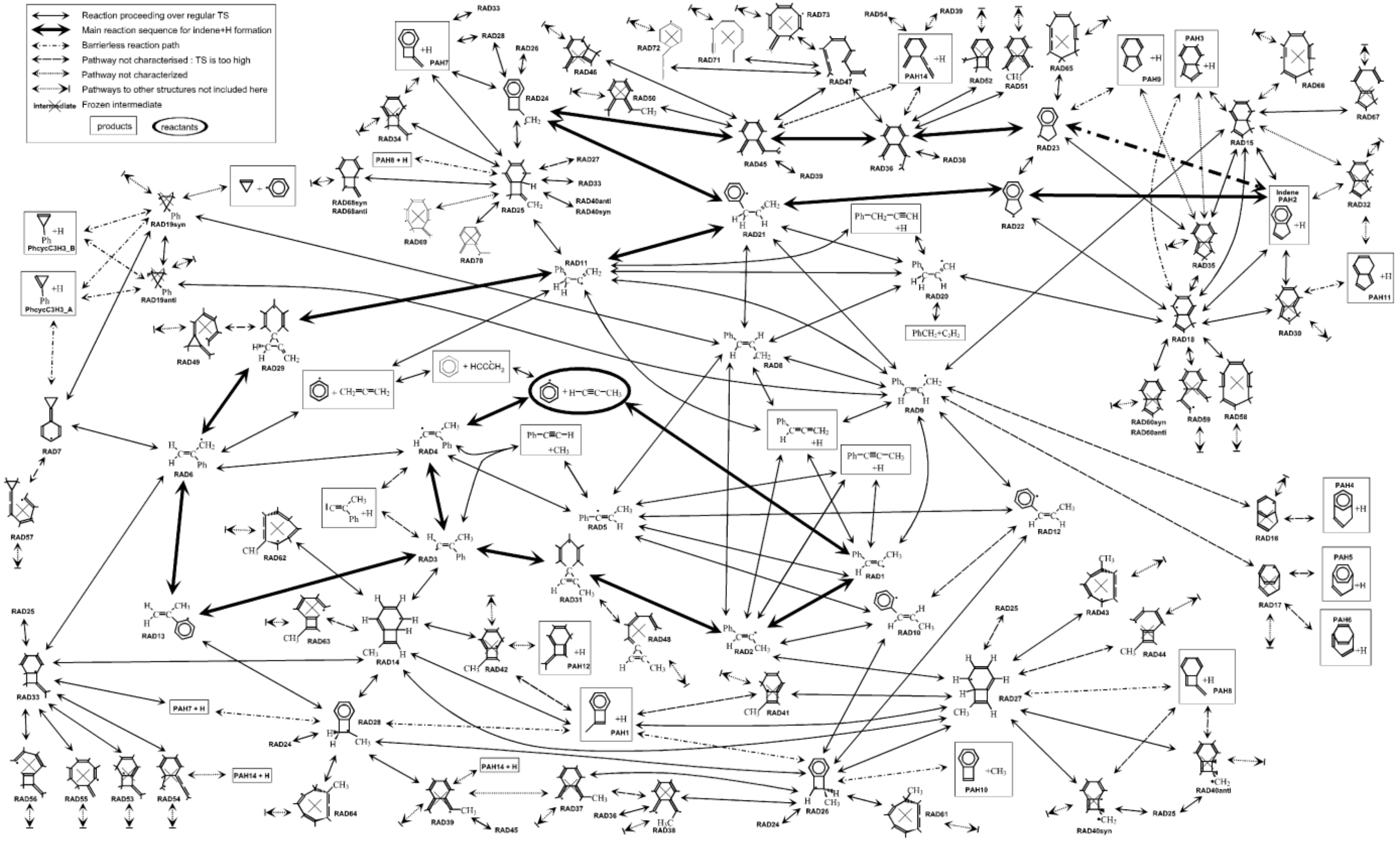
Typical Arrhenius representation



Framework

Plausible kinetic model is probably wrong…

…after discussion with your gr. leader…new plausible model…



Framework

Chemometric will support you for:

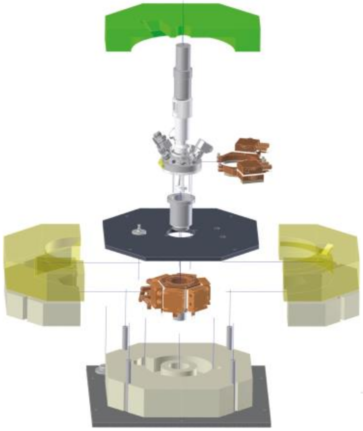
- How to design/select the reactor/instrumentation?
- How to do the experimental design?
- How to “find” the correct kinetic model?
- What to “fit” and how ?
- How to determine the rate constant?
- How to “fit” if “C” can not be isolated and unknown?
- What about baseline drift, shift, noise level?
- Finally, are my fitted parameters “correct” to which extends?
- And we have to be quick to do all the above tasks

1) Experimental setup

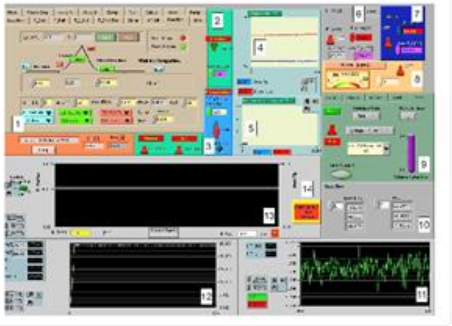
Parameters

- Custom designed and versatile
- Adapt lot of probes?
- Time dependent acquisition
- In-situ
- Well controlled
- Use of (SOPs, standard operation protocols)
- Fully automatized
- Fast data acquisition
- Acquire as much data as possible (then only average)

Reactor and control



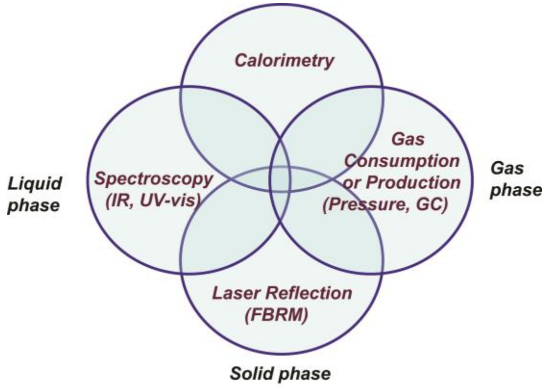
CAD



LabView

Instrumentation

- Adapt an orthogonal instrumental methodology



Calorimetry

Spectroscopy (IR, UV-vis)

Gas Consumption or Production (Pressure, GC)

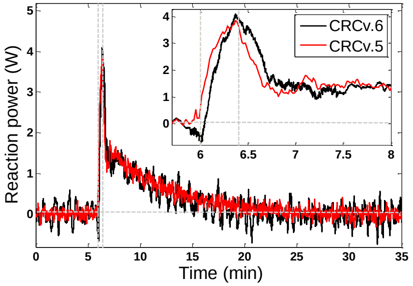
Laser Reflection (FBRM)

Liquid phase

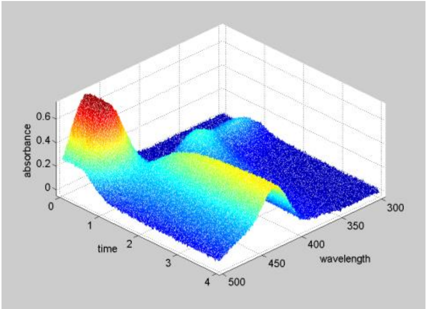
Gas phase

Solid phase

- Prefer multivariate signal to univariate signal
- Prefer integrated signals (spectroscopy) to differential signals (calorimetry)



Calorimetry

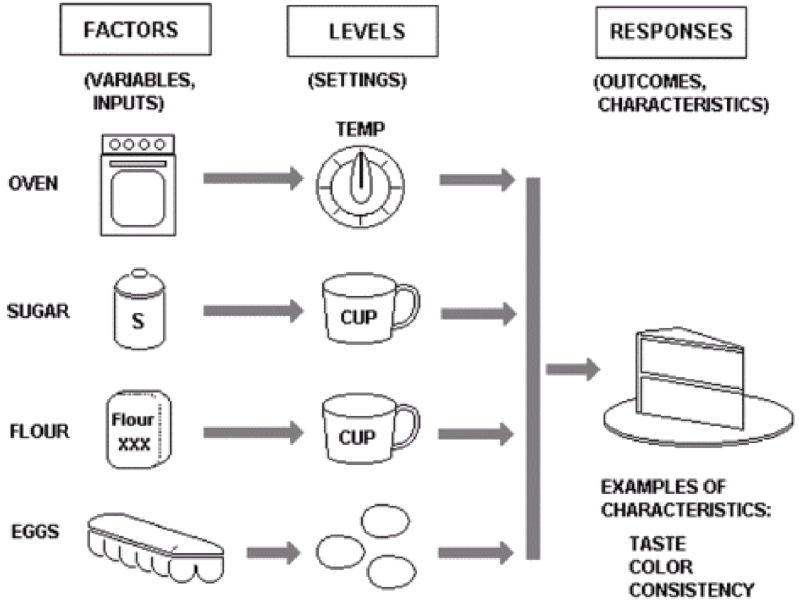


Spectroscopy

1) Design of Experiments (DoE)

Why a DoE?

- ➢ Used to reduce the design costs
- ➢ Speeds up the process design
- ➢ Reduce late engineering design
- ➢ Reduce raw and product material
- ➢ Reduce experimental complexity
- ➢ Reduce data analysis complexity
- ➢ Improve the overall robustness of data analysis
- ➢ Improve the certainty on the fitted parameters



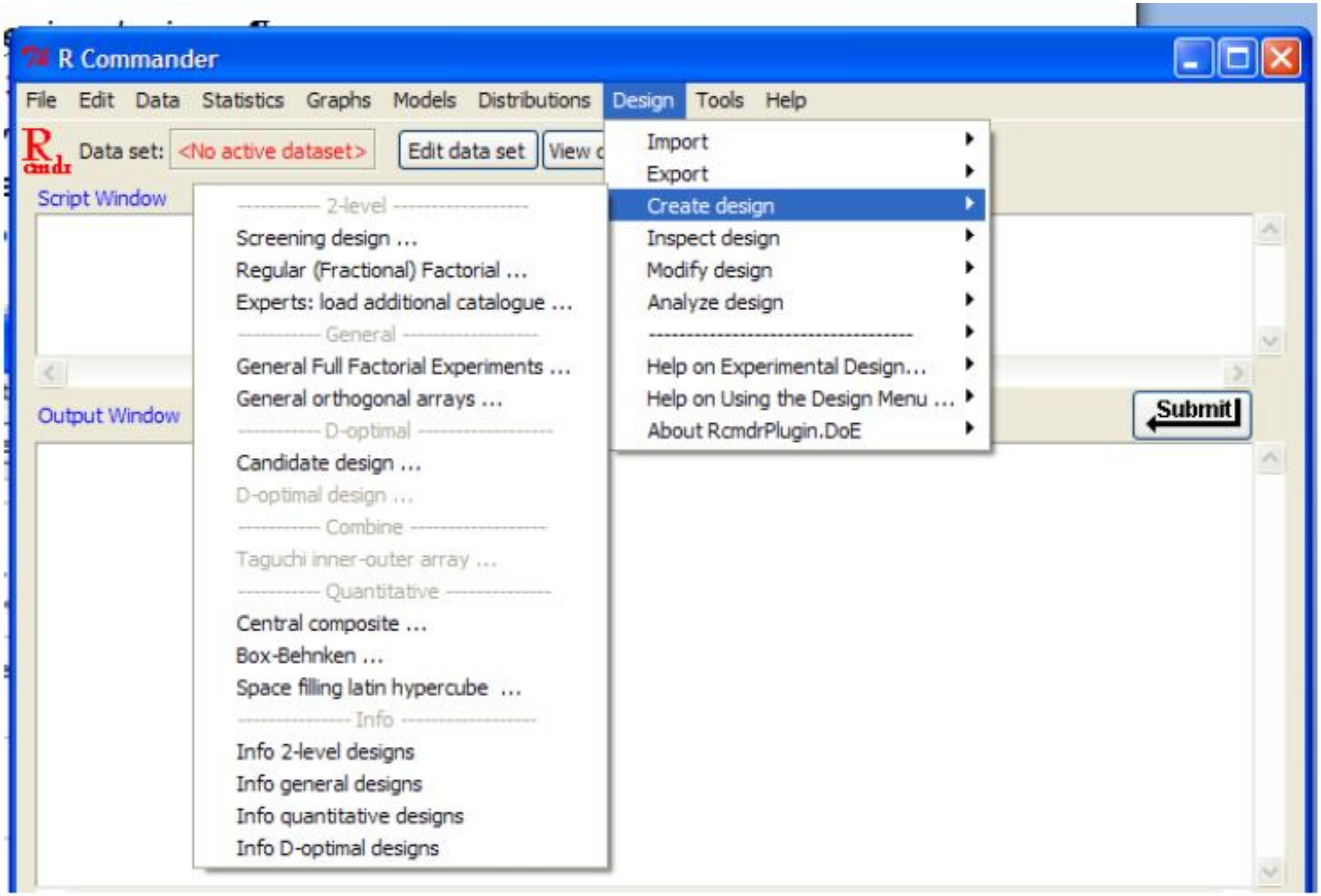
Experimental purpose

- ➢ Comparing alternative
  - ➢ Egg from different supplier gives the same cake?
- ➢ Significant inputs (Factors)
  - ➢ Sort the relevance of factors: is the amount of sugar more important than the amount of flower for the cake taste?
- ➢ Optimal process output (Response)
  - ➢ What are the “best parameters” to have the tastiest cake.
- ➢ Reducing variability (sensitivity analysis)
  - ➢ Which parameters can change slightly without changing the cake taste?
- ➢ Target an output
  - ➢ Which parameters for, best taste, best color, best consistency?
- ➢ Balancing tradeoffs
  - ➢ How to define optimum parameters to have the best taste at the best price

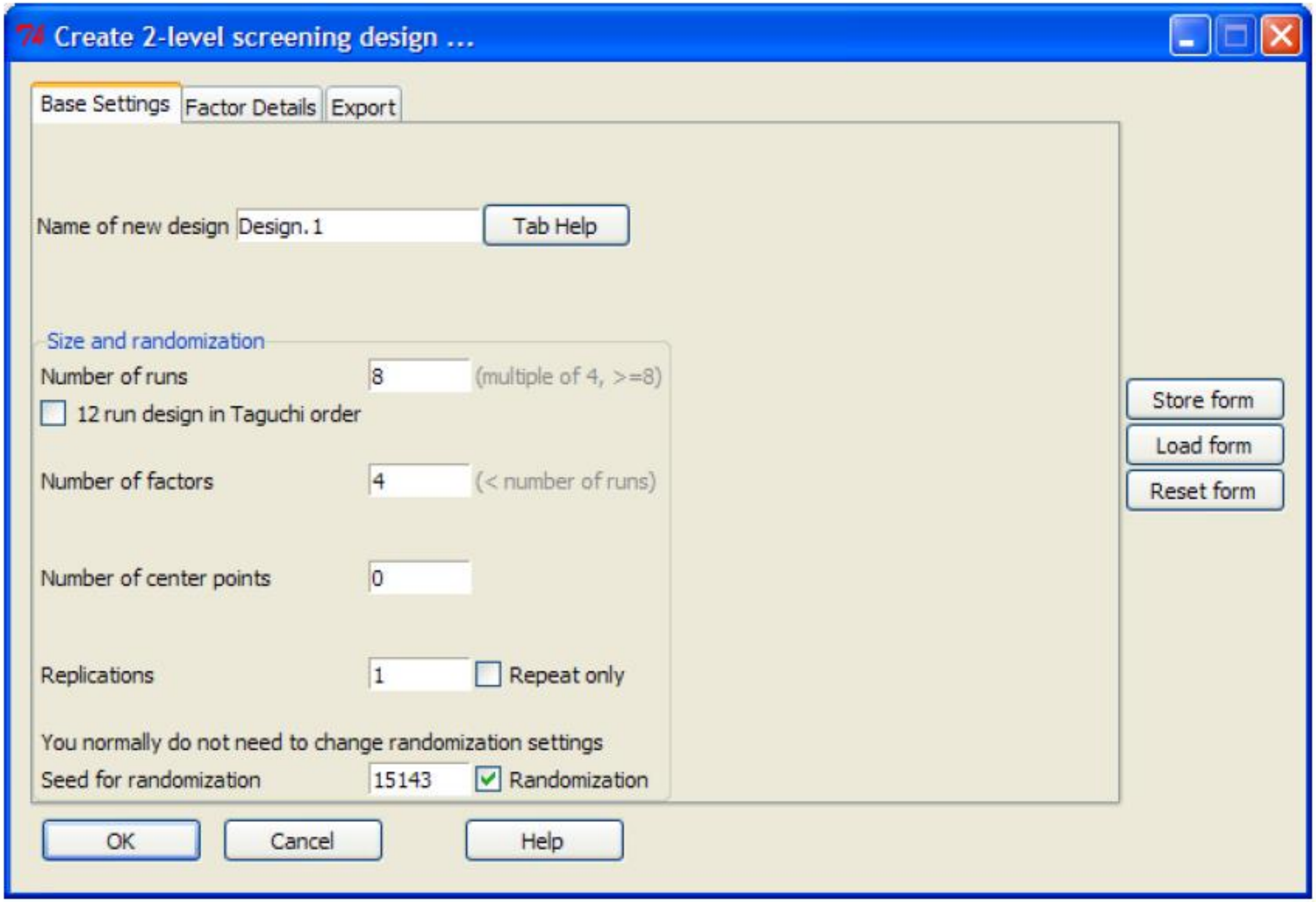
Nice DoE tutorial and “package “ :
I advise the R package RcmdrPlugin.DoE

1) Design of Experiments (DoE)

Tutorial for designing experiments using the R package RcmdrPlugin.DoE



Tutorial for designing experiments using the R package RcmdrPlugin.DoE



Create 2-level screening design ...

Base Settings | Factor Details | Export

Name of new design: Design.1

Size and randomization

Number of runs: 8 (multiple of 4, >=8)

□ 12 run design in Taguchi order

Number of factors: 4 (< number of runs)

Number of center points: 0

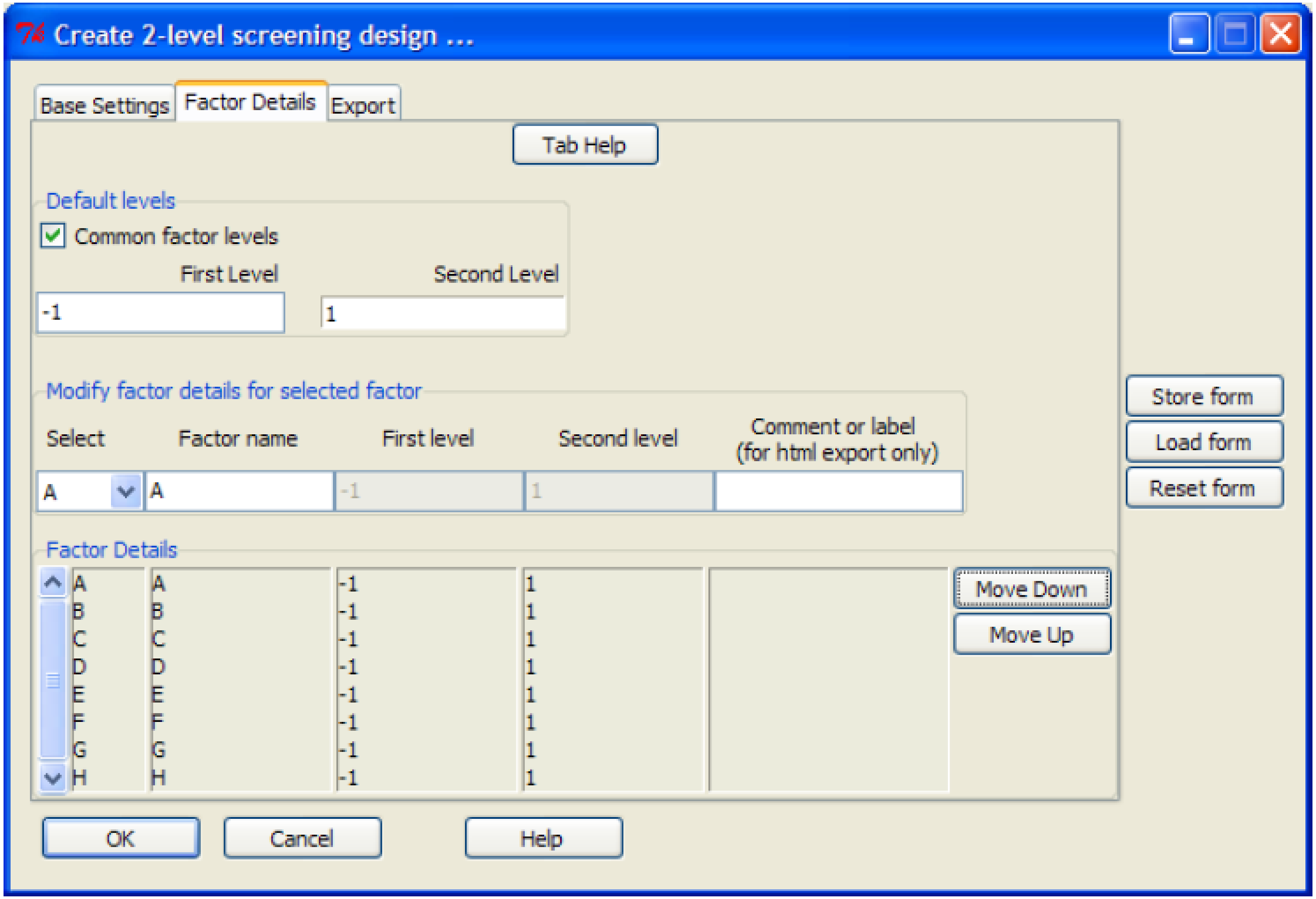
Replications: 1 □ Repeat only

You normally do not need to change randomization settings

Seed for randomization: 15143 ☑ Randomization

1) Design of Experiments (DoE)

Tutorial for designing experiments using the R package RcmdrPlugin.DoE



Create 2-level screening design …

Default levels

☑ Common factor levels

First Level: -1

Second Level: 1

Modify factor details for selected factor

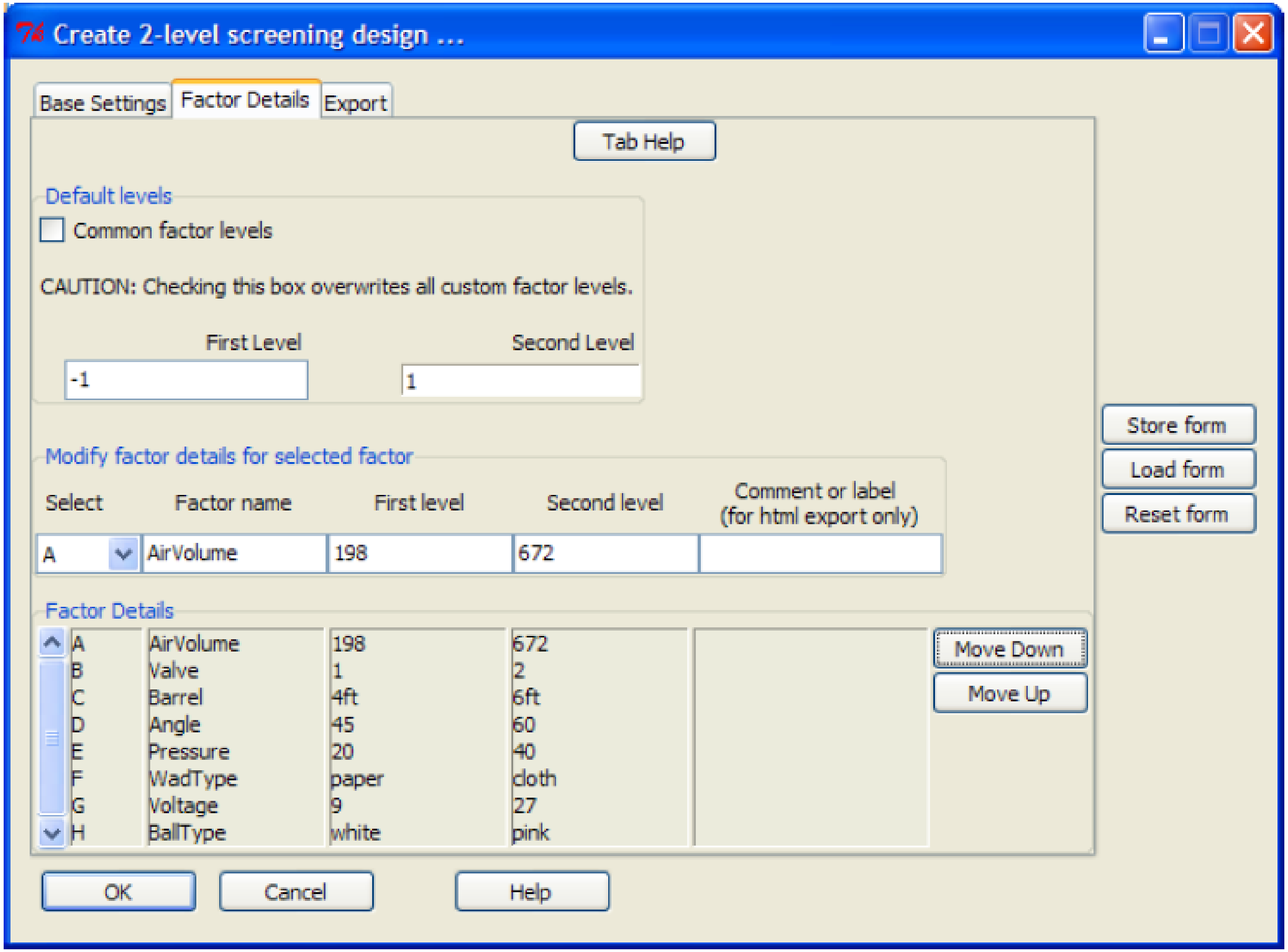
Select	Factor name	First level	Second level	Comment or label (for html export only)
A	A	-1	1	

Factor Details

A	A	-1	1	
B	B	-1	1	
C	C	-1	1	
D	D	-1	1	
E	E	-1	1	
F	F	-1	1	
G	G	-1	1	
H	H	-1	1	

1) Design of Experiments (DoE)

Tutorial for designing experiments using the R package RcmdrPlugin.DoE



Create 2-level screening design ...

Default levels

□ Common factor levels

CAUTION: Checking this box overwrites all custom factor levels.

First Level: -1

Second Level: 1

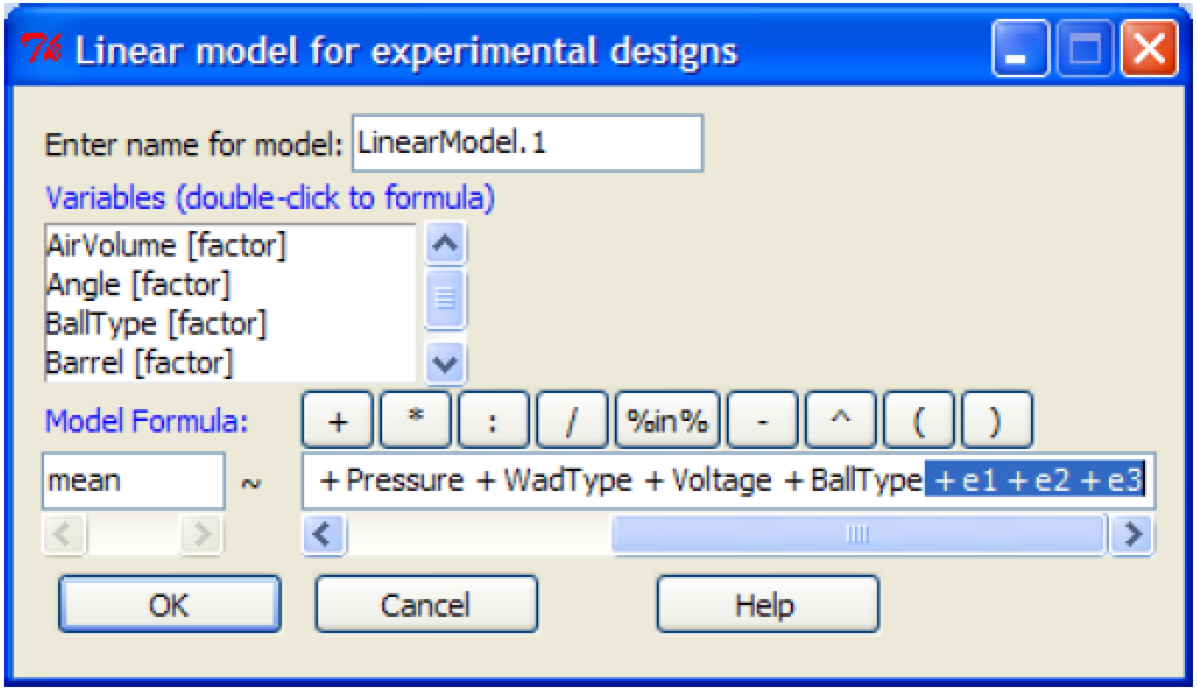
Modify factor details for selected factor

Select	Factor name	First level	Second level	Comment or label (for html export only)
A	AirVolume	198	672	

Factor Details

A	AirVolume	198	672	
B	Valve	1	2	
C	Barrel	4ft	6ft	
D	Angle	45	60	
E	Pressure	20	40	
F	WadType	paper	cloth	
G	Voltage	9	27	
H	BallType	white	pink	

1) Design of Experiments (DoE)



Linear model for experimental designs

Enter name for model: LinearModel.1

Variables (double-click to formula)

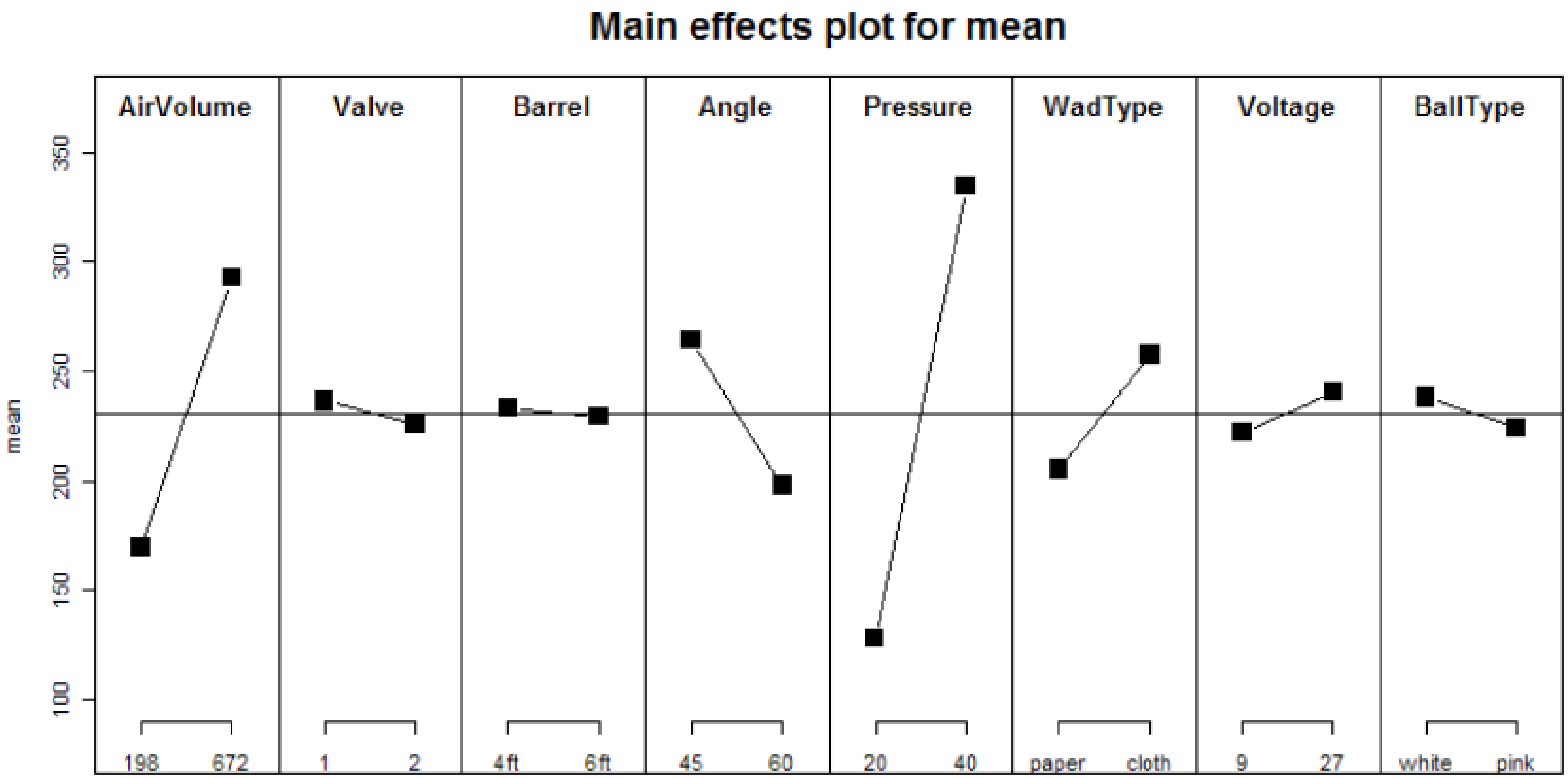
AirVolume [factor]
Angle [factor]
BallType [factor]
Barrel [factor]

Model Formula:

mean ~ + Pressure + WadType + Voltage + BallType + e1 + e2 + e3

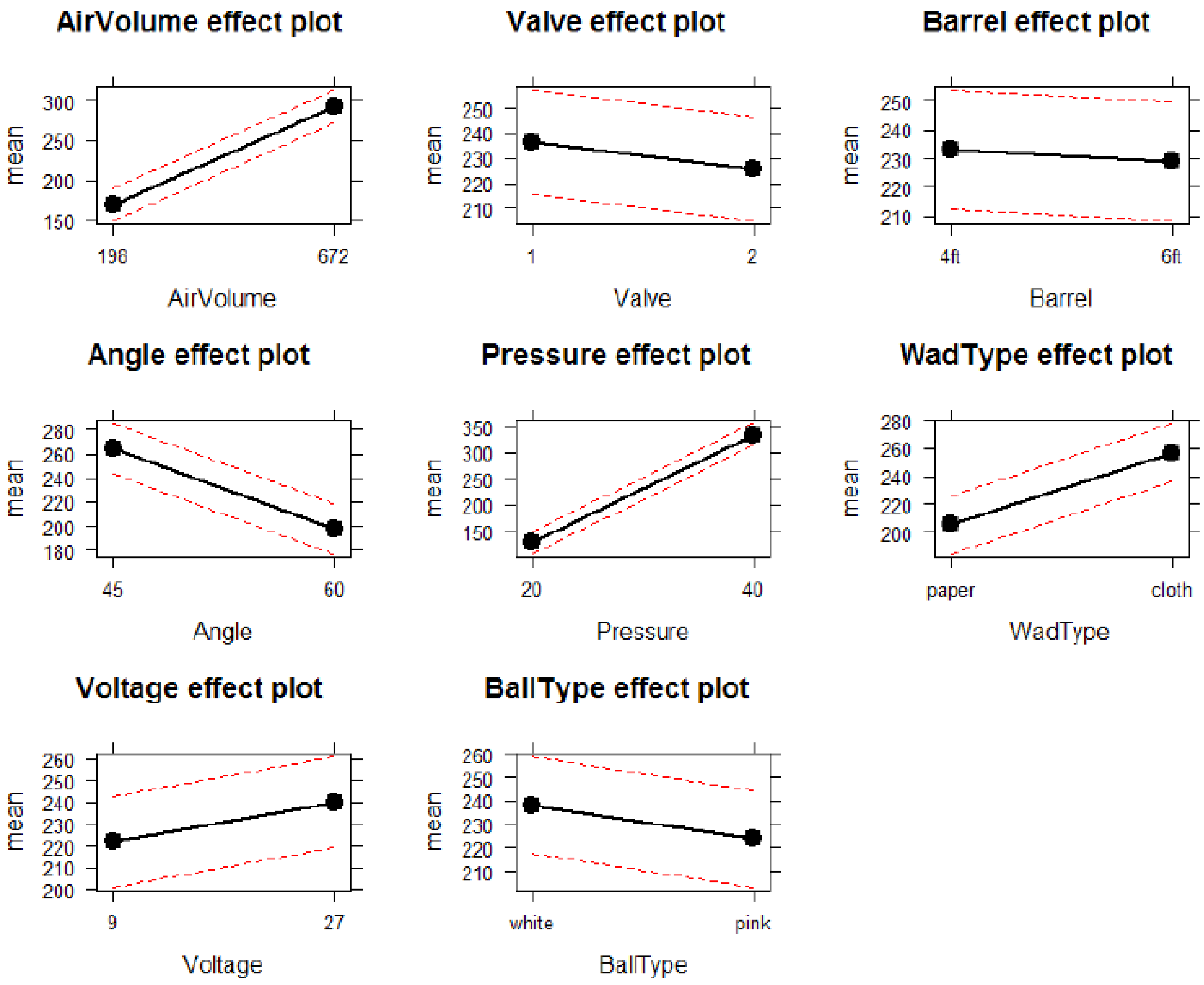
http://www1.beuth-hochschule.de/FB_II/reports/welcome.htm

Tutorial for designing experiments using the R package RcmdrPlugin.DoE



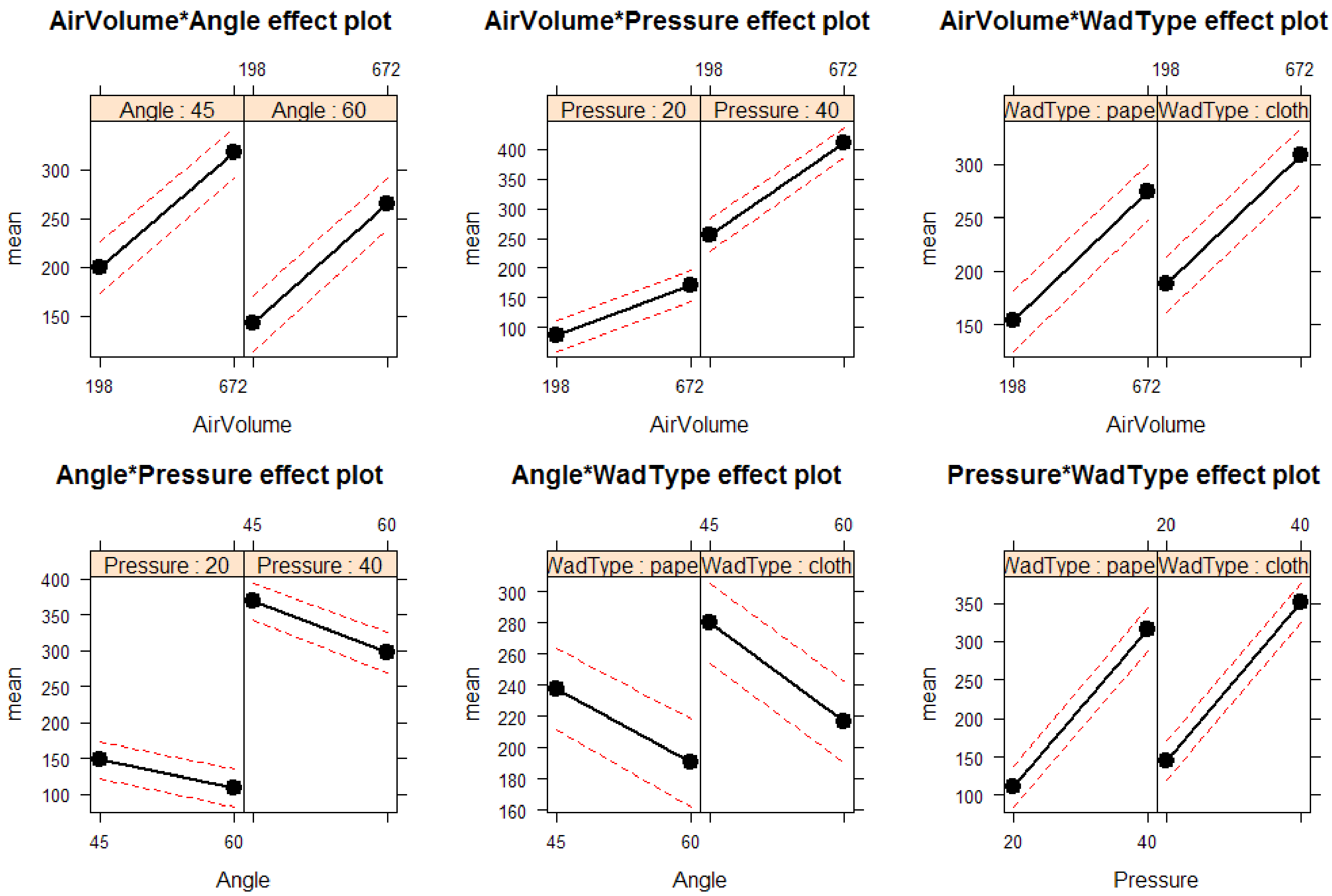
1) Design of Experiments (DoE)

Effects plots with model uncertainty



1) Design of Experiments (DoE)

Interaction plots



1) Design of Experiments (DoE)

- Comparing alternative
  - Egg from different supplier gives the same cake?
- Significant inputs (Factors)
  - Sort the relevance of factors: is the amount of sugar more important than the amount of flower for the cake taste?
- Optimal process output (Response)
  - What are the “best parameters” to have the tastiest cake.
- Reducing variability (sensitivity analysis)
  - Which parameters can change slightly without changing the cake taste?
- Target an output
  - Which parameters for, best taste, best color, best consistency?
- Balancing tradeoffs
  - How to define optimum parameters to have the best taste at the best price

Table of content

1) Introduction / Framework

2) Instrumentation / Design of Experiment DoE

3) Kinetic soft modelling methods
(FA, EFA, MCR-ALS)

4) Kinetic hard modelling
Kinetic fitting of spectroscopic data with non-linear regression

3) Table of content: Kinetic soft modelling methods

3a) Soft modelling versus hard modelling

3b) Absorption spectroscopy

- Beer’s law in elegant matrix notation (**Y** = **C** × **A**)
- Non-unique factorisation of **Y** / rotational ambiguity

3c) Principal Component Analysis (**PCA**)

- Abstract Factor analysis (**AFA**) by Singular Value Decomposition (**SVD**)
- Chemical rank of the measurement matrix
- The number of absorbing species

3d) Evolving Factor Analysis (**EFA**)

- Evolutionary rank analysis by repeated SVD of sub matrices of **Y**
- The ‘Appearance’ & ‘Disappearance’ of absorbing species

3e) Multivariate Curve Resolution by Alternating Least-Squares (**MCR-ALS**)

- Model-free iterative decomposition of **Y** = **C** × **A** + **R**
- Ideas, principles, limitations

3a) Soft modelling versus hard modelling

Factor model

PCA, EFA, MCR-ALS

Soft

Number of species
Approx. conc. & spectra

Lambert-Beer

Kinetic model

$$\upsilon = k \cdot c_{\mathrm{BuOH}} \cdot c_{\mathrm{AA}} \cdot c_{\mathrm{TMG}}$$

$$\frac{dc_{\mathrm{BuOH}}}{dt} = -\upsilon + \frac{f_{in}}{V} \cdot (c_{in,\mathrm{BuOH}} - c_{\mathrm{BuOH}})$$

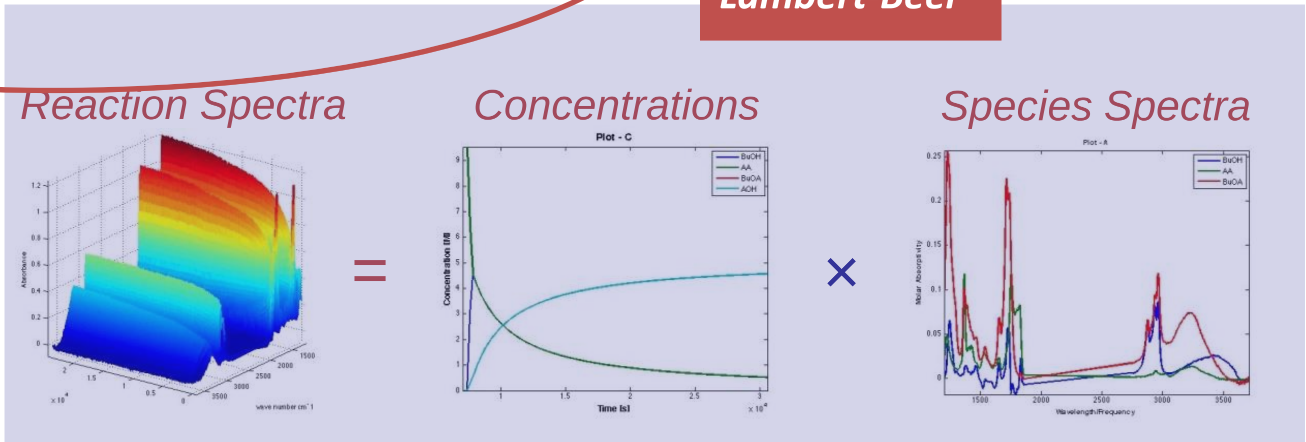
$$\frac{dc_{\mathrm{AA}}}{dt} = -\upsilon - \frac{f_{in}}{V} \cdot c_{\mathrm{AA}}$$

$$\frac{dc_{\mathrm{BuOA}}}{dt} = \upsilon - \frac{f_{in}}{V} \cdot c_{\mathrm{BuOA}}$$

$$\frac{dc_{\mathrm{AH}}}{dt} = \upsilon - \frac{f_{in}}{V} \cdot c_{\mathrm{AH}}$$

$$\frac{dc_{\mathrm{TMG}}}{dt} = -\frac{f_{in}}{V} \cdot c_{\mathrm{TMG}}$$

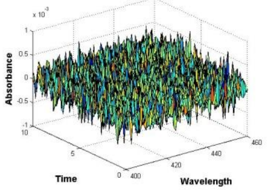
$$\frac{dV}{dt} = f_{in}$$



Hard

Rate constants
Activation energies

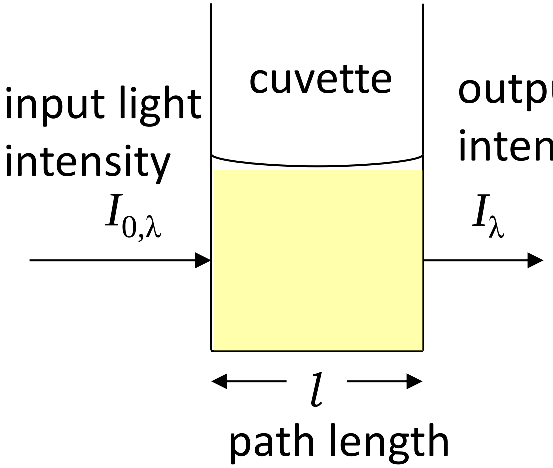
+ Residuals



Soft modelling: first insight into the experimental data

3b) Beer's law: Linear relationship between spectral absorbance and species concentrations

Absorption of UV-vis and IR light:


input light intensity $I_{0,\lambda}$

cuvette

output light intensity I_λ

l

path length

absorbance y at wavelength λ:

$$y_\lambda = -\log\left({}^{I}\!/\!_{I_0} \right)_\lambda$$

- absorbance signal y_λ [no unit] is linearly dependent on the concentrations c_s $[\mathrm{molL}^{-1}]$ of $s=1\ldots n_s$ absorbing species, the corresponding coefficients are the molar absorptivities $a_{s,\lambda}$ $[\mathrm{Lmol}^{-1}\mathrm{cm}^{-1}]$ and form the pure species spectra

Beer's Law:

$$y_\lambda = \left(c_1 a_{1,\lambda} + \ldots + c_s a_{s,\lambda} + \ldots + c_{n_s} a_{n_s,\lambda} \right) \times l$$

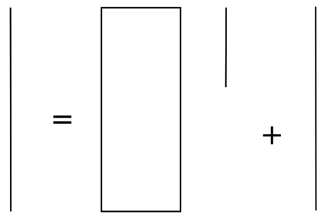
$$= \sum_{s=1}^{n_s} c_s a_{s,\lambda} \times l$$

Often:
path length l =1cm

3b) Nomenclature

- But it is time to introduce some useful conventions for a scalar/vector/matrix notation:
  - matrices are given in **boldface capital** characters, e.g. $\mathbf{Y}$
  - vectors are given in **boldface lower case** characters, e.g. $\mathbf{y}$, this includes row or column vectors of a matrix $\mathbf{Y}$, e.g. $\mathbf{y}_{2,:}$ (2nd row of $\mathbf{Y}$), or $\mathbf{y}_{:,3}$ (3rd column of $\mathbf{Y}$)
  - scalars are given in *italic* characters, this includes the elements of a vector $\mathbf{y}$ or matrix $\mathbf{Y}$, e.g. y_i or $y_{i,j}$

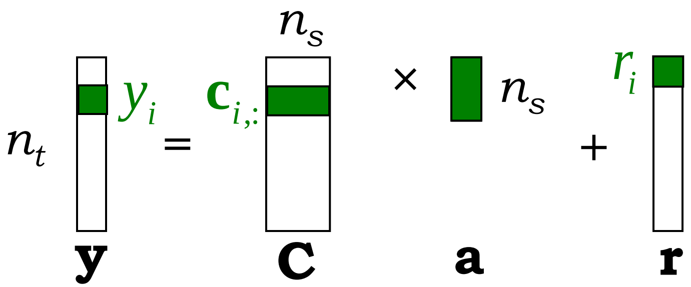
- In some cases, it is very illustrative to write vector/matrix equations in a "line/box" notation, e.g.



$$\mathbf{y} = \mathbf{F} \times \mathbf{a} + \mathbf{r}$$

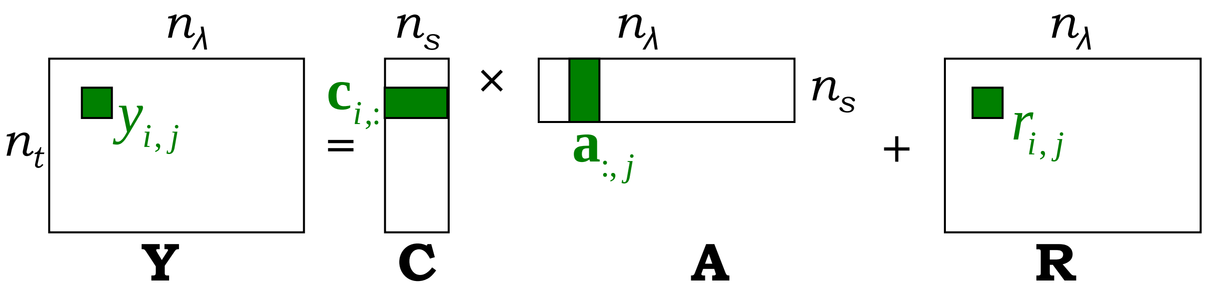
3b) Beer's law: Linear relationship between spectral absorbance and species concentrations

- For time and/or wavelength resolved kinetic data, Beer's law can be written in elegant vector (single wavelength) or matrix notation (multi wavelengths):



at $i=1\ldots n_t$ times

$$y_i = \sum_{s=1}^{n_s} c_{i,s} a_s + r_i$$

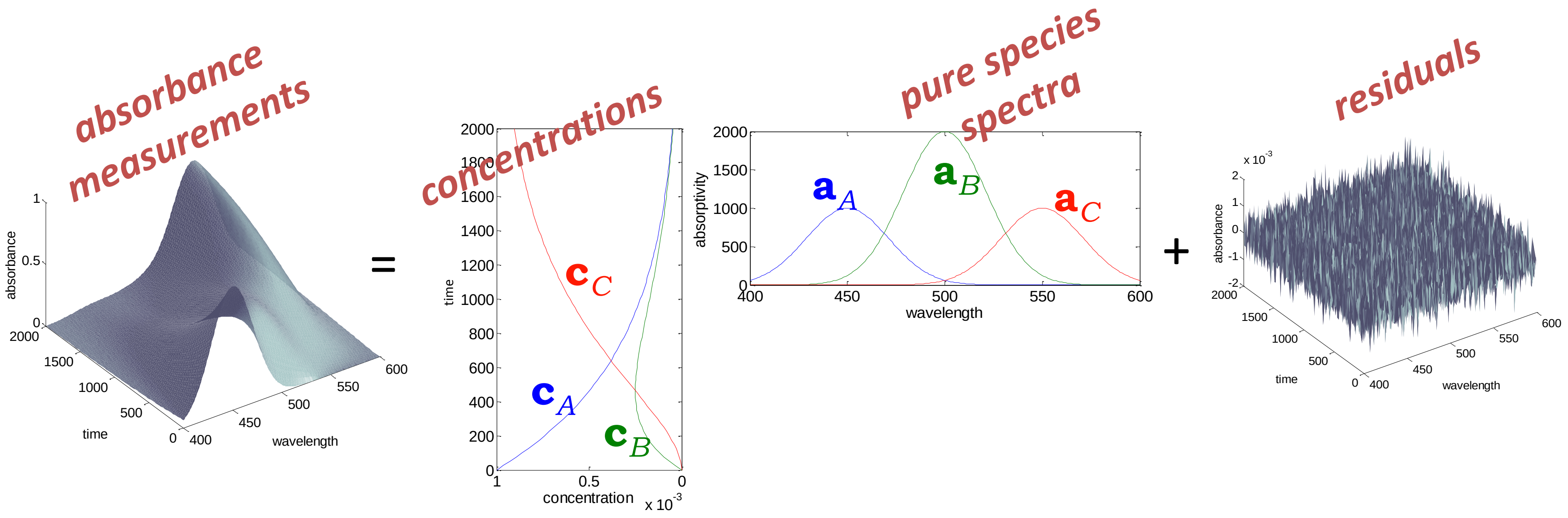


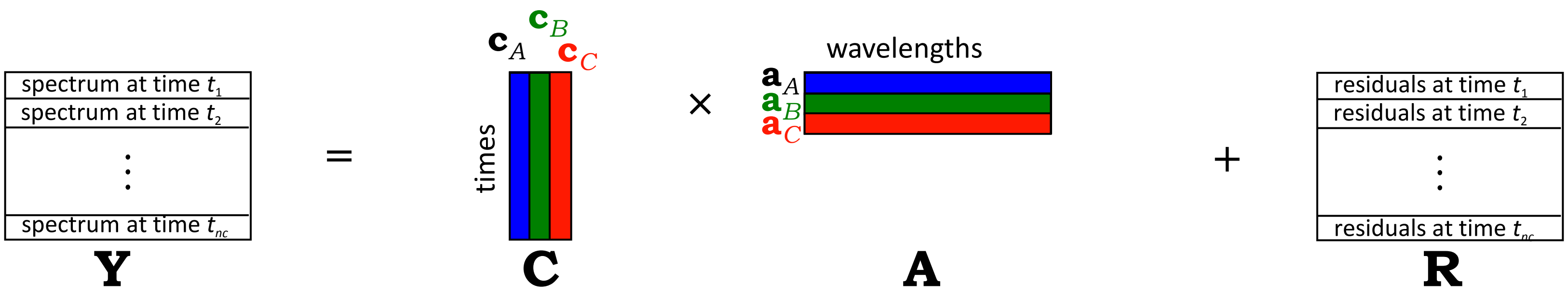
at $j=1\ldots n_\lambda$ wavelengths

$$y_{i,j} = \sum_{s=1}^{n_s} c_{i,s} a_{s,j} + r_{i,j}$$

3b) Beer's law: Linear relationship between spectral absorbance and species concentrations

for example: a reaction *A*→*B*→*C*





spectrum at time t_1

spectrum at time t_2

⋮

spectrum at time t_{nc}

$\mathbf{Y}$

=

$\mathbf{c}_A$ $\mathbf{c}_B$ $\mathbf{c}_C$

times

$\mathbf{C}$

×

wavelengths

$\mathbf{a}_A$

$\mathbf{a}_B$

$\mathbf{a}_C$

$\mathbf{A}$

+

residuals at time t_1

residuals at time t_2

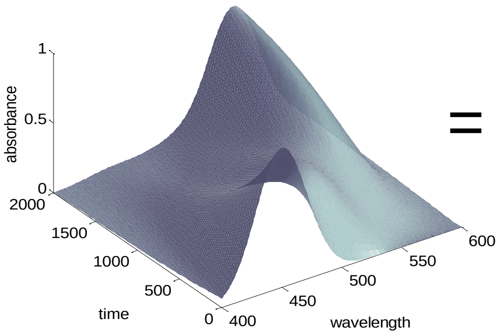
⋮

residuals at time t_{nc}

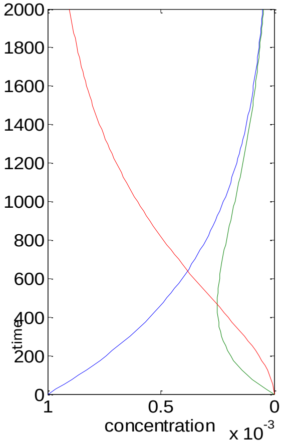
$\mathbf{R}$

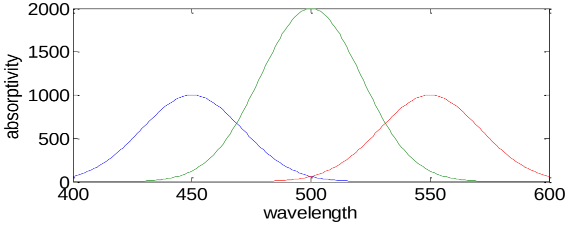
3b) Beer's law: Linear relationship between spectral absorbance and species concentrations

e.g. A→B→C

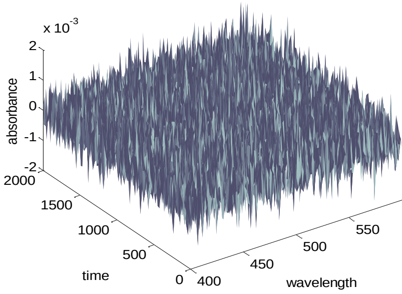


=





+



?

Y = **C** × **A** + **R**

Goal: Find concentration profiles **C** and species spectra **A** such that the residuals **R**=**Y**-**CA** become small only using a 'soft model', i.e. by linear factorisation

Problem: Factorisation is not unique (rotational ambiguity)

3c) Soft modelling classes

- By using appropriate ‘soft’ restrictions on **C** and **A**, e.g. non-negativity, windows of existence, closure, unimodality, known spectra, the number of possible solutions can be reduced, sometimes can even lead to a unique solution for **C** & **A**
- There are 2 major classes

1) Factor Analysis (AFA) based

$$\mathbf{Y} = \bar{\mathbf{U}}\bar{\mathbf{S}}\bar{\mathbf{V}} = \bar{\mathbf{U}}\mathbf{T}\mathbf{T}^{-1}\bar{\mathbf{S}}\bar{\mathbf{V}} = \mathbf{C}\mathbf{A}$$

$$\text{Find } \mathbf{T} \text{ such that}$$

$$\mathbf{C} = \bar{\mathbf{U}}\mathbf{T}, \text{ and } \mathbf{A} = \mathbf{T}^{-1}\bar{\mathbf{S}}\bar{\mathbf{V}}$$

2) Alternating Least Squares (ALS) based

$$\text{Start from some guessed } \mathbf{C},$$

$$\text{then recalculate } \mathbf{A} \text{ and } \mathbf{C} \text{ until satisfied:}$$

$$\mathbf{A} = (\mathbf{C}^{t}\mathbf{C})^{-1}\mathbf{C}^{t}\mathbf{Y} = \mathbf{C}^{+}\mathbf{Y}$$

$$\mathbf{C} = \mathbf{Y}\mathbf{A}^{t}(\mathbf{A}\mathbf{A}^{t})^{-1} = \mathbf{Y}\mathbf{A}^{+}$$

3) Table of content: Kinetic soft modelling methods

3a) Soft modelling versus hard modelling

3b) Absorption spectroscopy

- Beer's law in elegant matrix notation (**Y** = **C** × **A**)
- Non-unique factorisation of **Y** / rotational ambiguity

3c) Principal Component Analysis (**PCA**)

- Abstract Factor analysis (**AFA**) by Singular Value Decomposition (**SVD**)
- Chemical rank of the measurement matrix
- The number of absorbing species

3d) Evolving Factor Analysis (**EFA**)

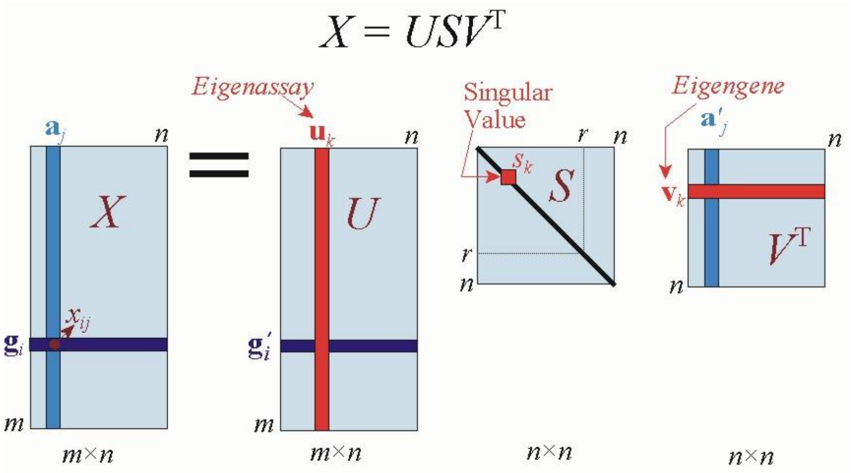
- Evolutionary rank analysis by repeated SVD of sub matrices of **Y**
- The 'Appearance' & 'Disappearance' of absorbing species

3e) Multivariate Curve Resolution by Alternating Least-Squares (**MCR-ALS**)

- Model-free iterative decomposition of **Y** = **C** × **A** + **R**
- Ideas, principles, limitations

3c) Principle component analysis

- One very well defined solution is the one received from Abstract Factor Analysis (AFA) using Singular Value Decomposition (SVD)



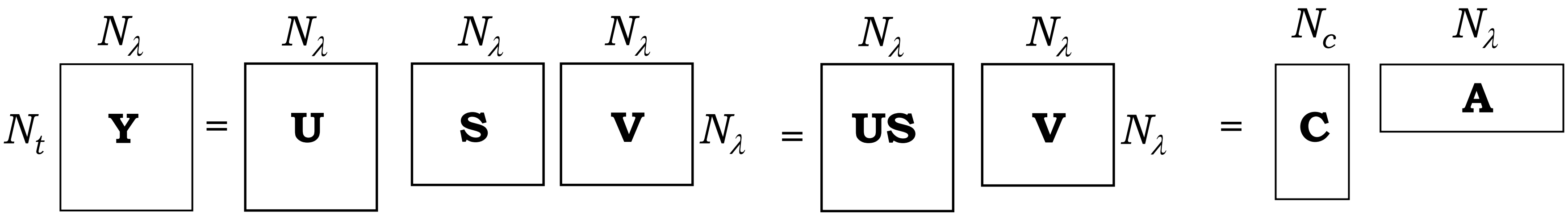
S is a diagonal matrix with the square root of their eigenvalues

columns of **U** (rows of **V**) are eigenvectors of **YY**^{t} (**Y**^{t}**Y**)

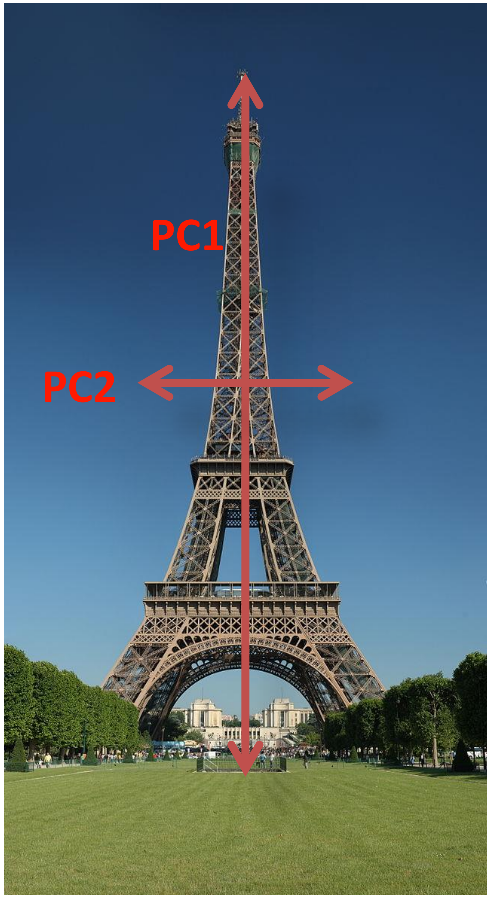
U and **V**^{t} are orthonormal

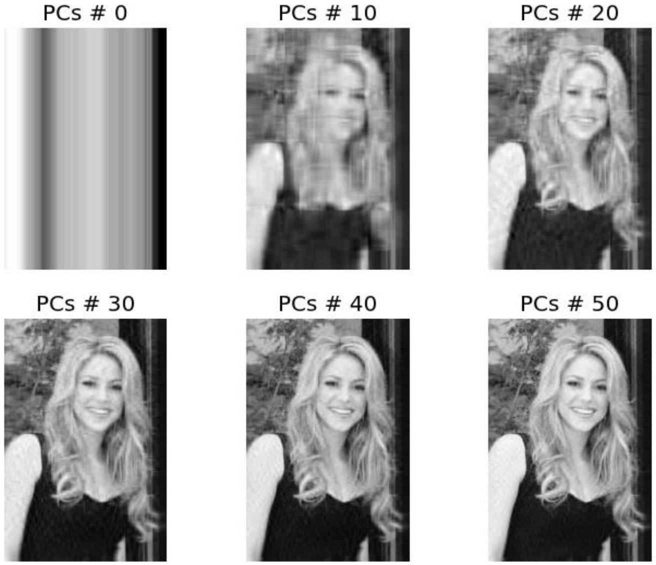
in Matlab:

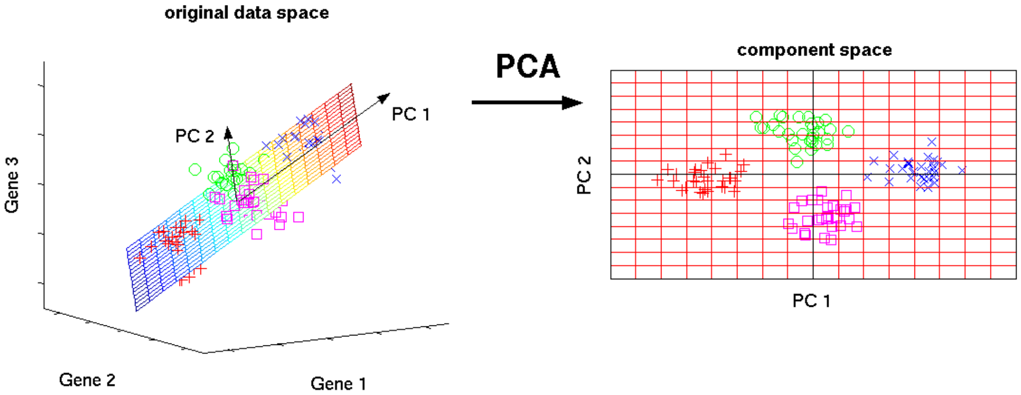
```
[U,S,Vt]=svd(Y,0);
```



3c) Principle component analysis



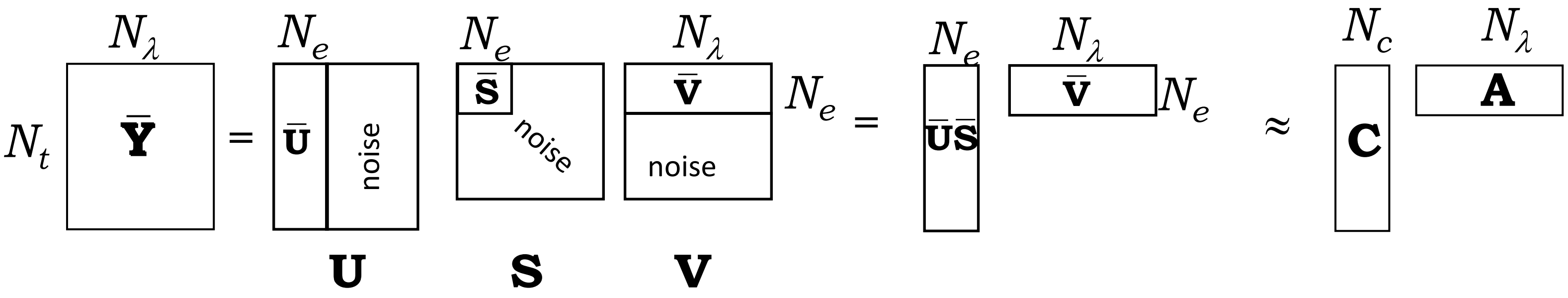




Indeed very efficient tool for data compression and noise filtration (orthogonal noise is removed

3c) Principle component analysis

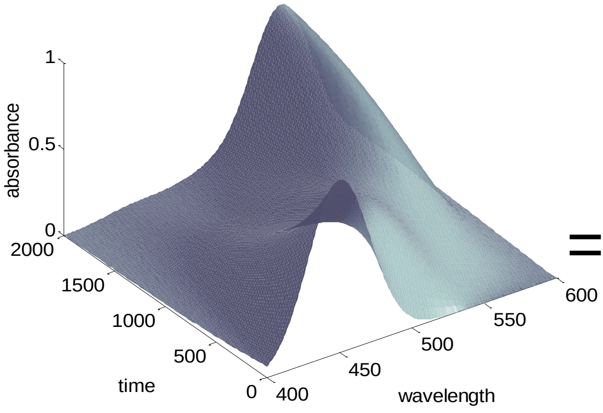
- Eigenvectors in **U** (columns) and **V** (rows) are arranged in decreasing order of magnitude of their corresponding singular values in **S**
- Many of them just represent ‘noise’ and can be neglected; the significant ‘factors’, the Principal Components, are retained in $\bar{\mathbf{U}}$ and $\bar{\mathbf{V}}$ and form ‘abstract’ concentration profiles and spectra
- The diagonal elements of $\bar{\mathbf{S}}$, the singular values, can be seen as normalisation coefficients for $\bar{\mathbf{U}}$ or $\bar{\mathbf{V}}$



- The number of significant singular(eigen) values and –vectors is the chemical rank of **Y** and a 1st estimate on the number of absorbing species

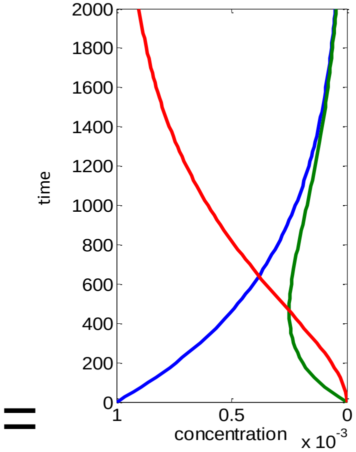
3c) Principle component analysis

e.g. A→B→C

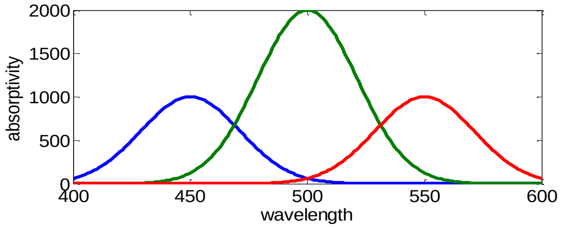


$\mathbf{Y}$

$=$

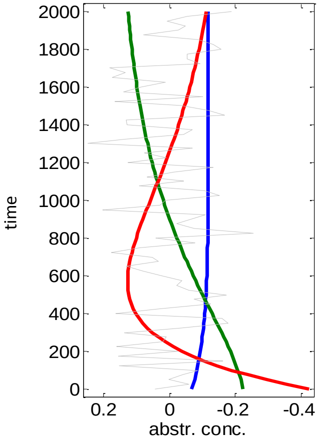


$\mathbf{C}$



$\mathbf{A}$

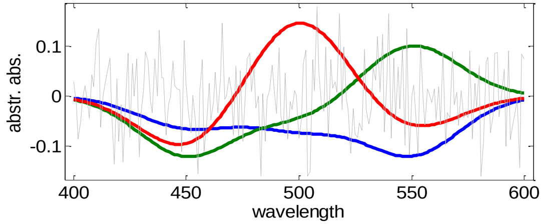
$+\ \mathbf{R}$



$\bar{\mathbf{U}}$

$$\begin{pmatrix} 43 & & 0 \\ & 22 & \\ 0 & & 6 \end{pmatrix} 10^{-2}$$

$\bar{\mathbf{S}}$



$\bar{\mathbf{V}}$

$+\ \mathbf{R}_{PCA}$

→ rank(**Y**) = 3

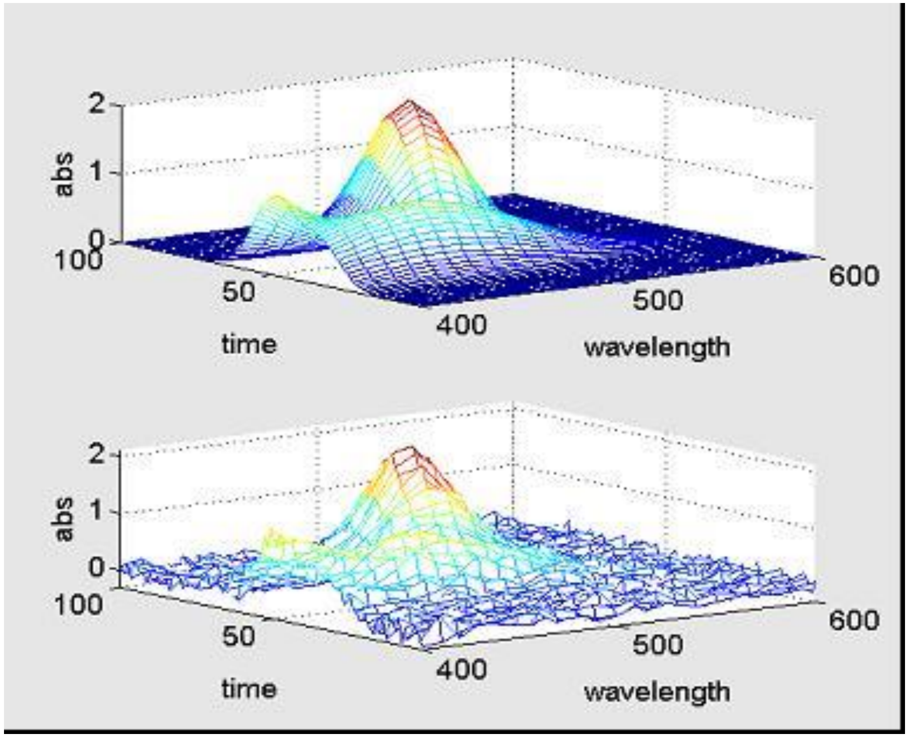
→ 3 absorbing species

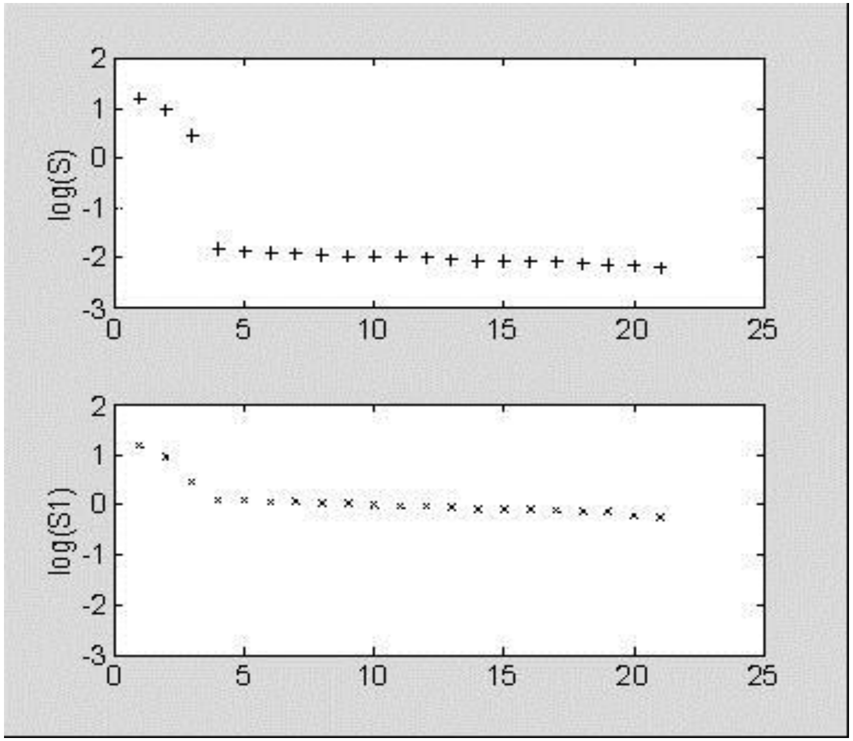
3c) Chemical rank – number of absorbing species

The **rank** of a matrix *A* is the size of the largest collection of linearly independent columns or rows of *A.*

Y $(N_t \times N_\lambda)$

log($s_{i,i}$) vs *i*





The noise level in the data matrix **Y** determines the drop in the magnitude from significant to insignificant singular values

3) Table of content: Kinetic soft modelling methods

3a) Soft modelling versus hard modelling

3b) Absorption spectroscopy

- Beer’s law in elegant matrix notation (**Y** = **C** × **A**)
- Non-unique factorisation of **Y** / rotational ambiguity

3c) Principal Component Analysis (**PCA**)

- Abstract Factor analysis (**AFA**) by Singular Value Decomposition (**SVD**)
- Chemical rank of the measurement matrix
- The number of absorbing species

3d) Evolving Factor Analysis (**EFA**)

- Evolutionary rank analysis by repeated SVD of sub matrices of **Y**
- The ‘Appearance’ & ‘Disappearance’ of absorbing species

3e) Multivariate Curve Resolution by Alternating Least-Squares (**MCR-ALS**)

- Model-free iterative decomposition of **Y** = **C** × **A** + **R**
- Ideas, principles, limitations

3d) EFA: Chemical rank – number of absorbing species

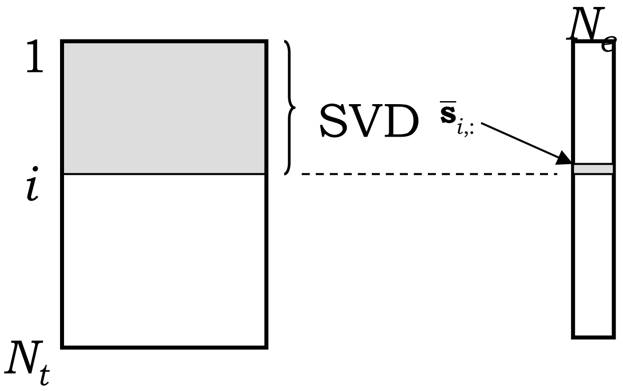
- the chemical rank of the spectral data matrix **Y** is determined by the number of its significant singular vectors
- the number of significant singular vectors of **Y** is determined by the number of linearly independent columns or rows in the matrix of pure species spectra (**A**) and corresponding concentration profiles (**C**)
- linear dependencies in **C** due to the kinetic model are common and sometimes difficult to predict (e.g. *A*+*B*→*C*)
- linear dependencies in **A** are less common

3d) Evolving Factor Analysis (EFA)

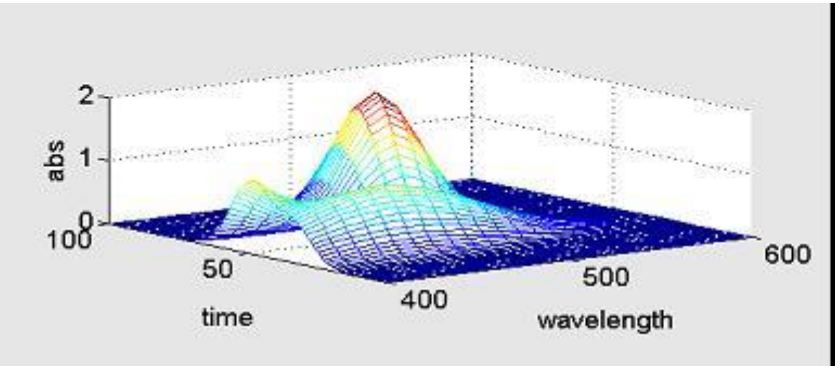
- sequential rank analysis of the data matrix along its time domain by repeated SVD
- can be performed in a forward and backward way
- indicates the rise of new singular vectors and thus gives an estimate for the appearance & disappearance of new absorbing species
- ideally designed to follow chromatography experiments
  - species appear & disappear sequentially
- capable of roughly following kinetic profiles
  - species can appear & dissappear simultaneously

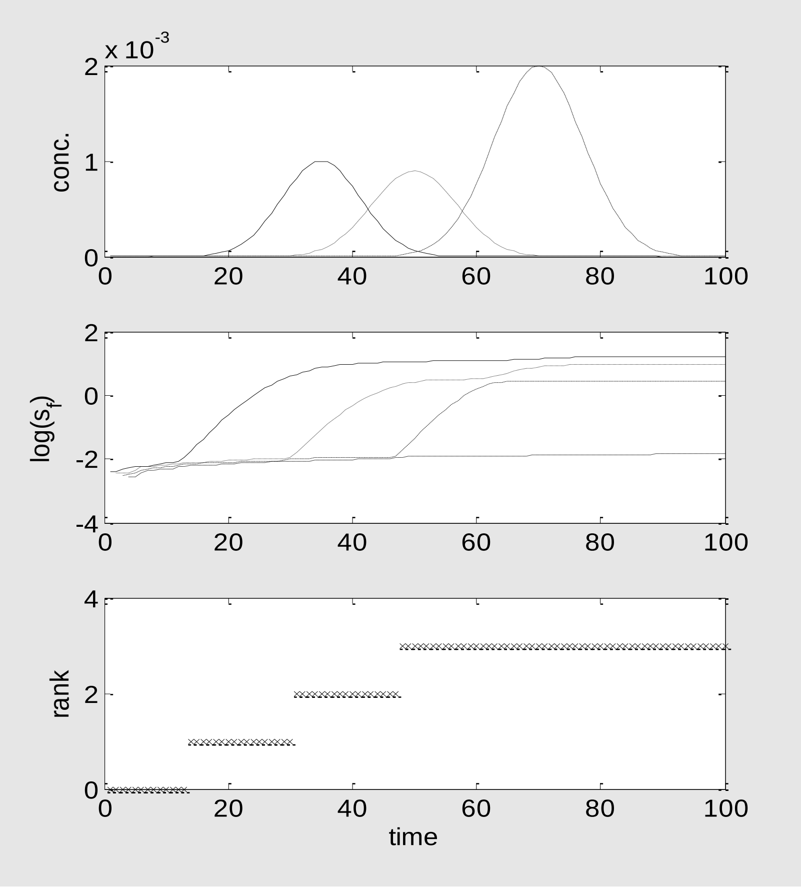
3d) Evolving Factor Analysis (EFA)

Forward EFA



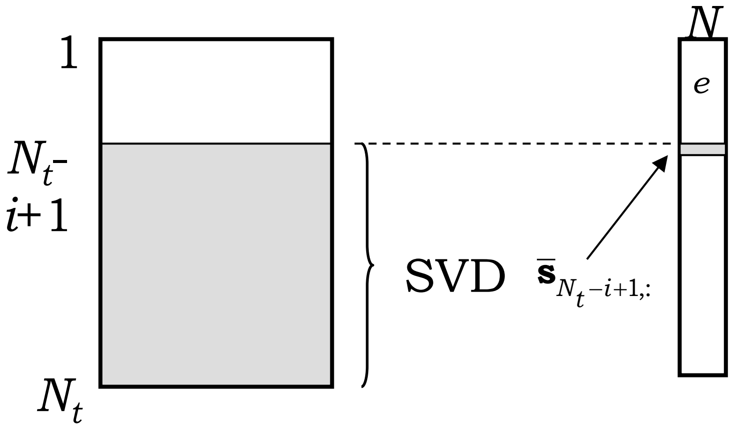
- Repeated rank analysis by SVD in forward direction
- The appearance of a new ‘species’ is indicated by a gradual rise of a new singular value





3d) Evolving Factor Analysis (EFA)

Backward EFA



1

N_t-

i+1

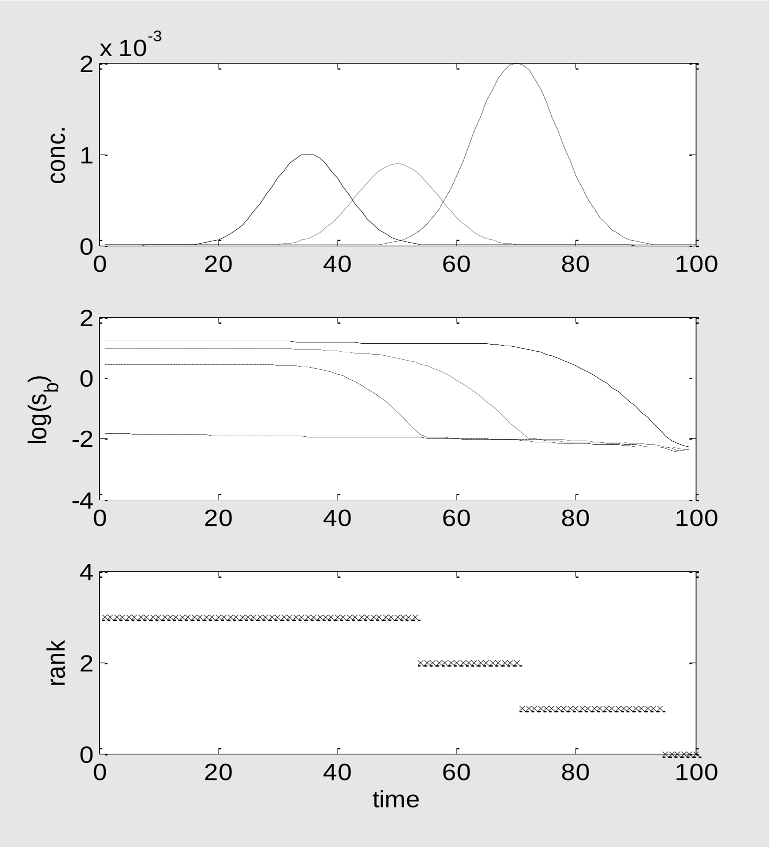
N_t

SVD

$\bar{\mathbf{s}}_{N_t-i+1,:}$

N

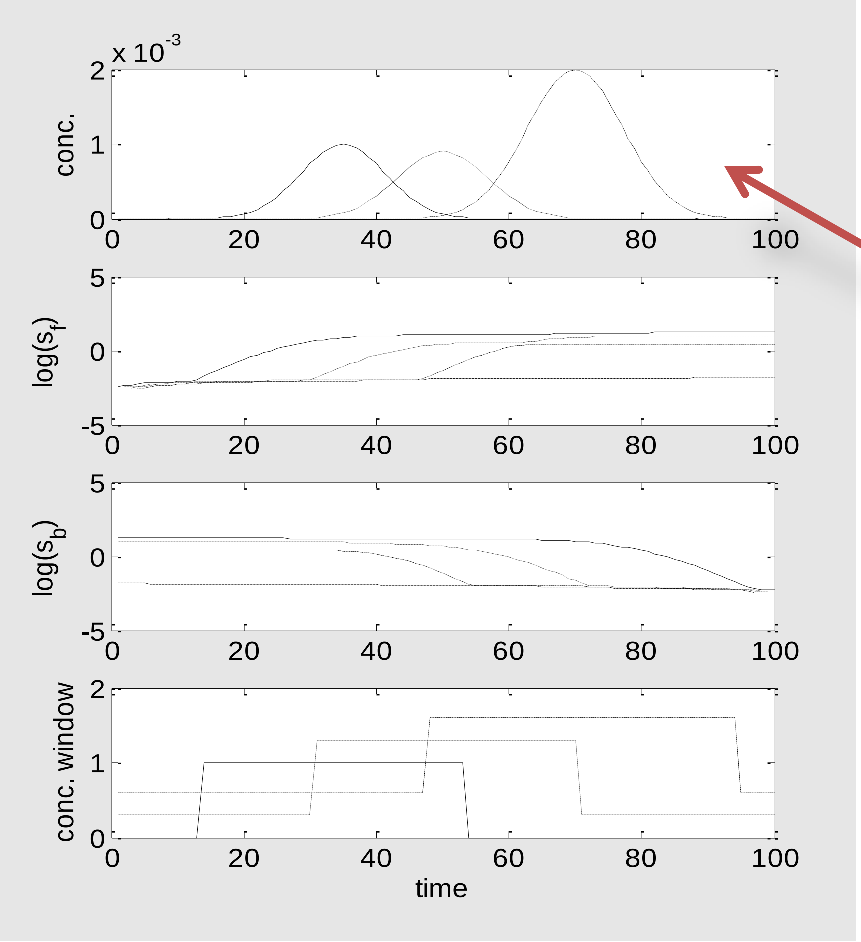
e



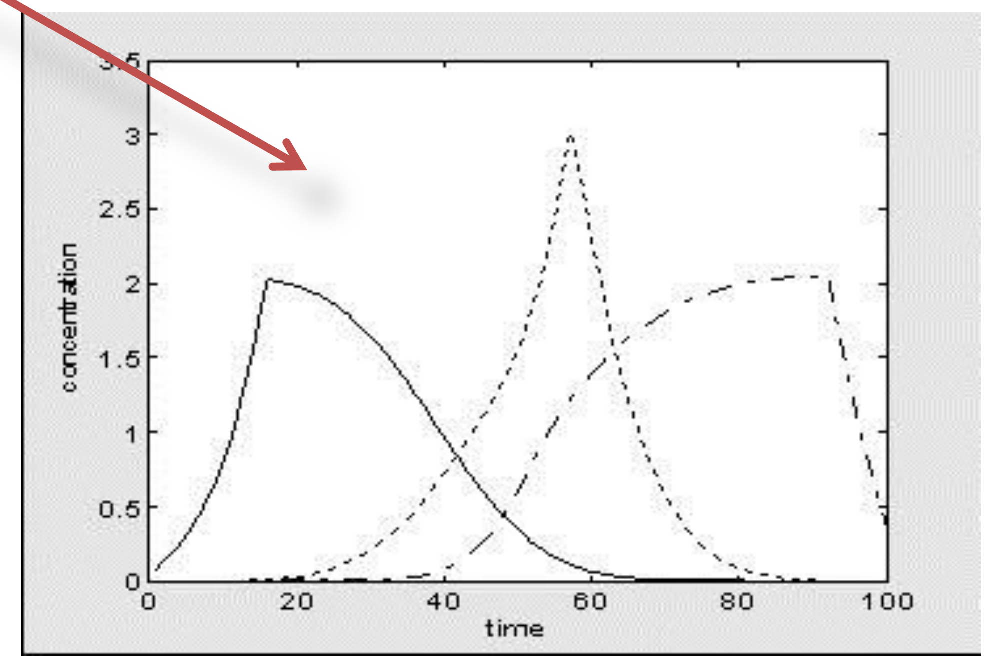
- Repeated rank analysis by SVD in backward direction
- A ‘disappearing species’ is indicated by a gradual rise of a new singular value

3d) Evolving Factor Analysis (EFA)

Forward and backward EFA



Combined forward/backward EFA results can be used as reasonable initial guesses of concentration profiles for subsequent iterative refinement e.g. by ALS



3) Table of content: Kinetic soft modelling methods

3a) Soft modelling versus hard modelling

3b) Absorption spectroscopy

- Beer’s law in elegant matrix notation ($\mathbf{Y} = \mathbf{C} \times \mathbf{A}$)
- Non-unique factorisation of **Y** / rotational ambiguity

3c) Principal Component Analysis (**PCA**)

- Abstract Factor analysis (**AFA**) by Singular Value Decomposition (**SVD**)
- Chemical rank of the measurement matrix
- The number of absorbing species

3d) Evolving Factor Analysis (**EFA**)

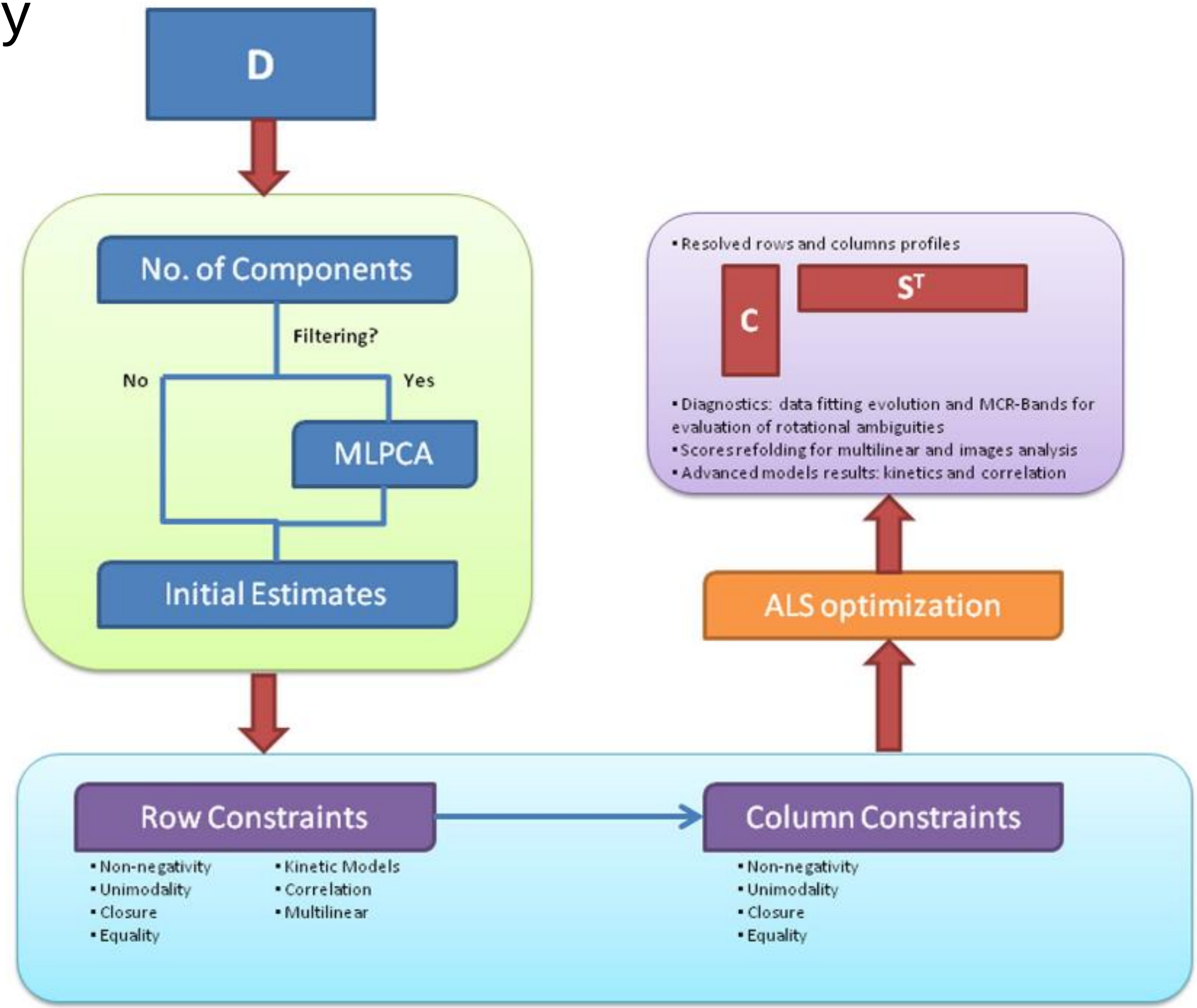
- Evolutionary rank analysis by repeated SVD of sub matrices of **Y**
- The ‘Appearance’ & ‘Disappearance’ of absorbing species

3e) Multivariate Curve Resolution by Alternating Least-Squares (**MCR-ALS**)

- Model-free iterative decomposition of $\mathbf{Y} = \mathbf{C} \times \mathbf{A} + \mathbf{R}$
- Ideas, principles, limitations

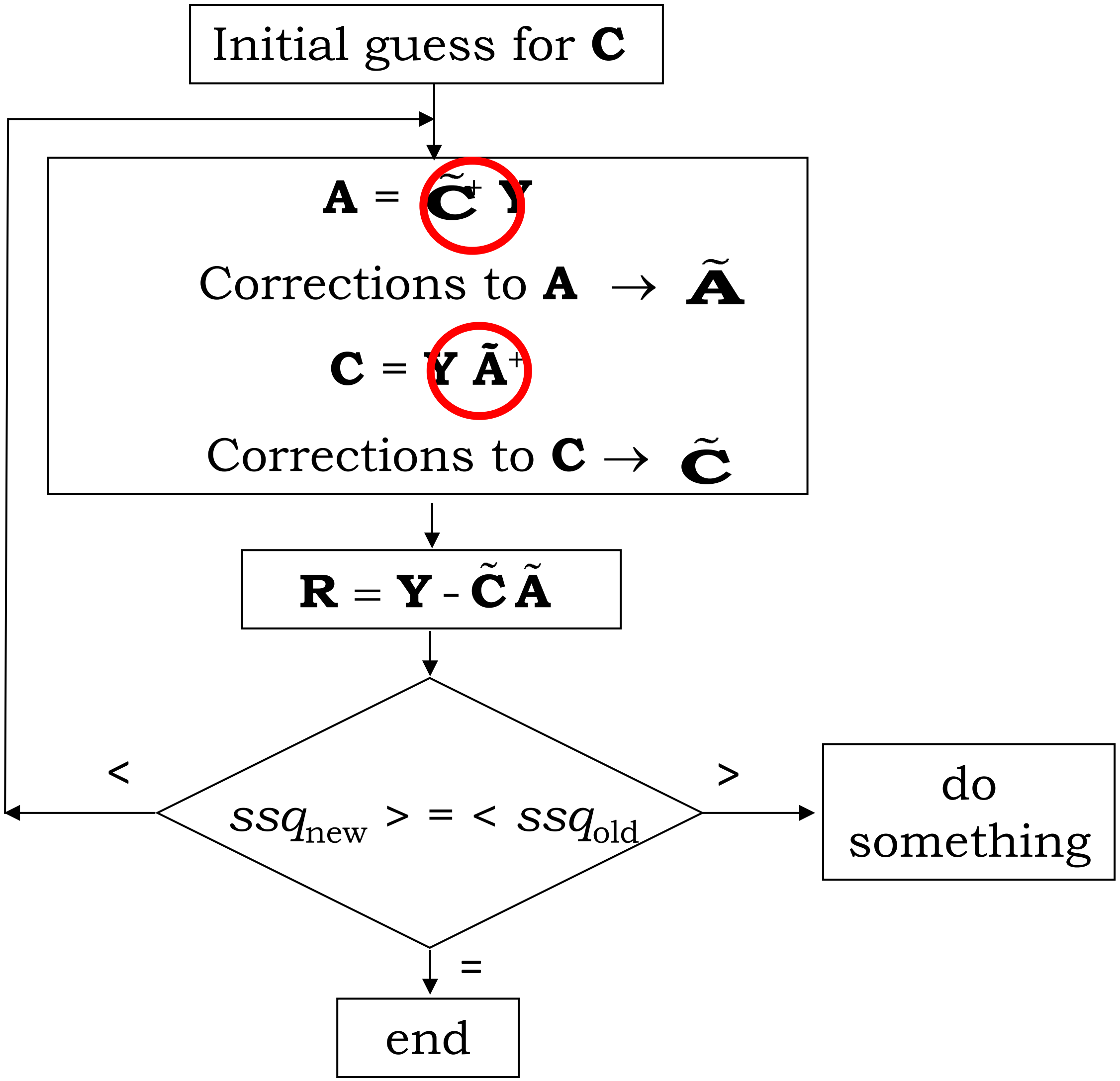
3e) Multivariate curve resolution by alternating least-squares (MCR-ALS)

- Conceptually very simple



http://www.mcrals.info/

3e) Multivariate curve resolution by alternating least-squares (MCR-ALS)



Initial guess for **C**

$\mathbf{A} = \tilde{\mathbf{C}}^{+} \mathbf{Y}$

Corrections to **A** → $\tilde{\mathbf{A}}$

$\mathbf{C} = \mathbf{Y} \tilde{\mathbf{A}}^{+}$

Corrections to **C** → $\tilde{\mathbf{C}}$

$\mathbf{R} = \mathbf{Y} - \tilde{\mathbf{C}} \tilde{\mathbf{A}}$

<

$ssq_{\mathrm{new}} > = < ssq_{\mathrm{old}}$

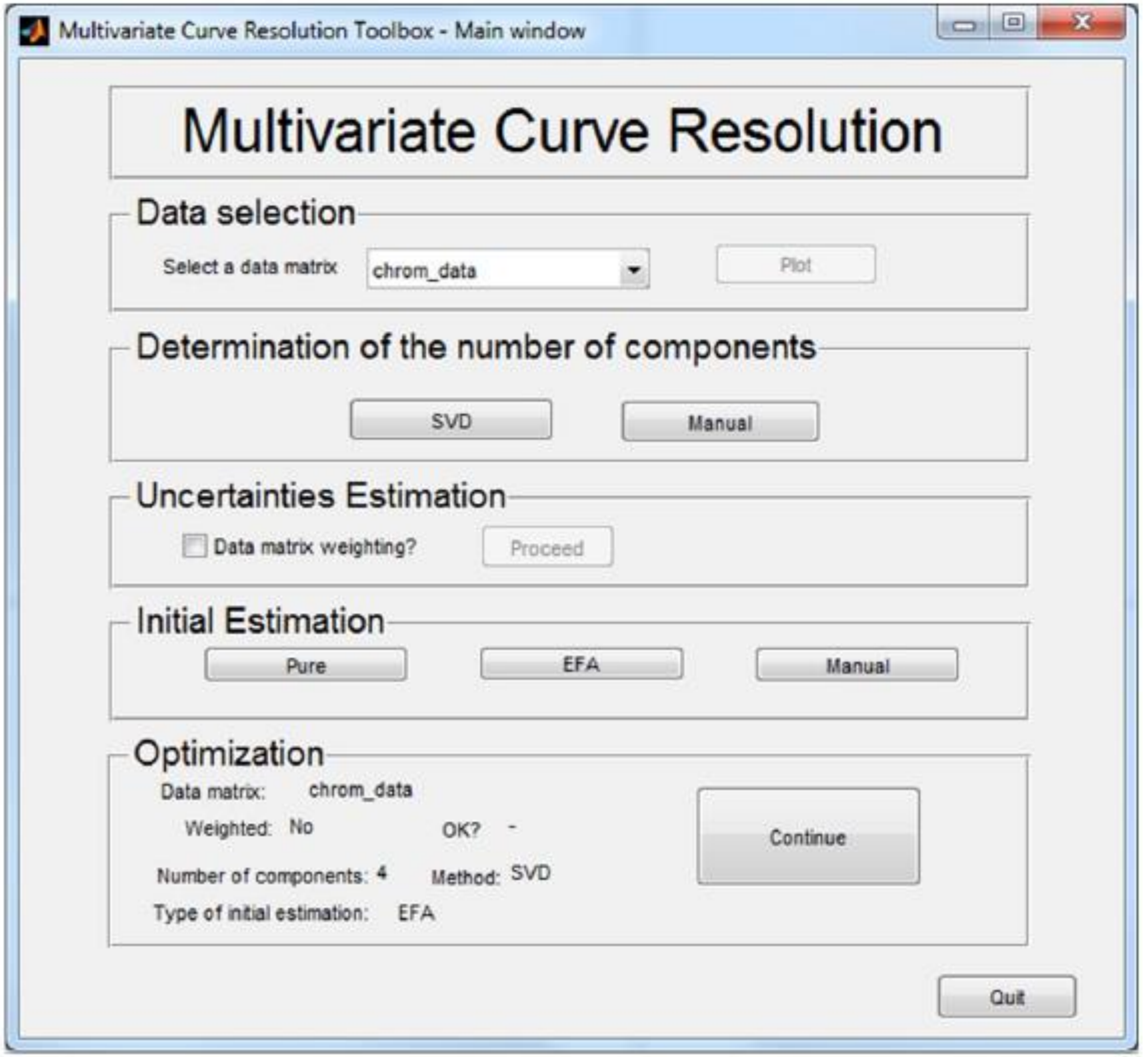
>

do something

=

end

3e) Multivariate curve resolution by alternating least-squares (MCR-ALS)



Multivariate Curve Resolution Toolbox - Main window

Multivariate Curve Resolution

Data selection

Select a data matrix: chrom_data

Determination of the number of components

Uncertainties Estimation

☐ Data matrix weighting?

Initial Estimation

Optimization

Data matrix: chrom_data

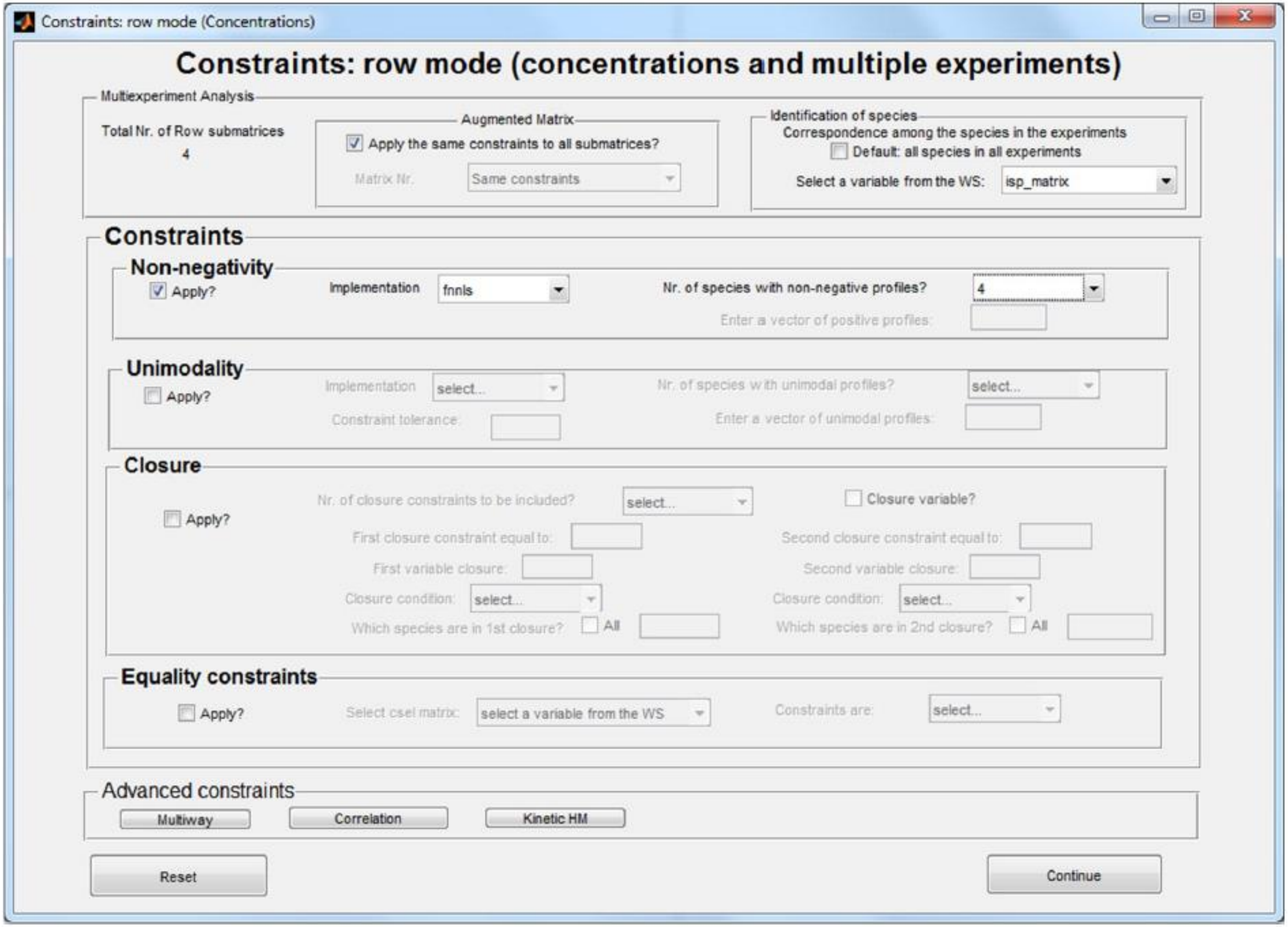
Weighted: No   OK? -

Number of components: 4   Method: SVD

Type of initial estimation: EFA

http://www.mcrals.info/

3e) Multivariate curve resolution by alternating least-squares (MCR-ALS)



Constraints: row mode (Concentrations)

Constraints: row mode (concentrations and multiple experiments)

Multiexperiment Analysis

Total Nr. of Row submatrices
4

Augmented Matrix
☑ Apply the same constraints to all submatrices?
Matrix Nr. Same constraints

Identification of species
Correspondence among the species in the experiments
□ Default: all species in all experiments
Select a variable from the WS: isp_matrix

Constraints

Non-negativity
☑ Apply?
Implementation: fnnls
Nr. of species with non-negative profiles? 4
Enter a vector of positive profiles:

Unimodality
□ Apply?
Implementation: select...
Nr. of species with unimodal profiles? select...
Constraint tolerance:
Enter a vector of unimodal profiles:

Closure
□ Apply?
Nr. of closure constraints to be included? select...
□ Closure variable?
First closure constraint equal to:
Second closure constraint equal to:
First variable closure:
Second variable closure:
Closure condition: select...
Closure condition: select...
Which species are in 1st closure? □ All
Which species are in 2nd closure? □ All

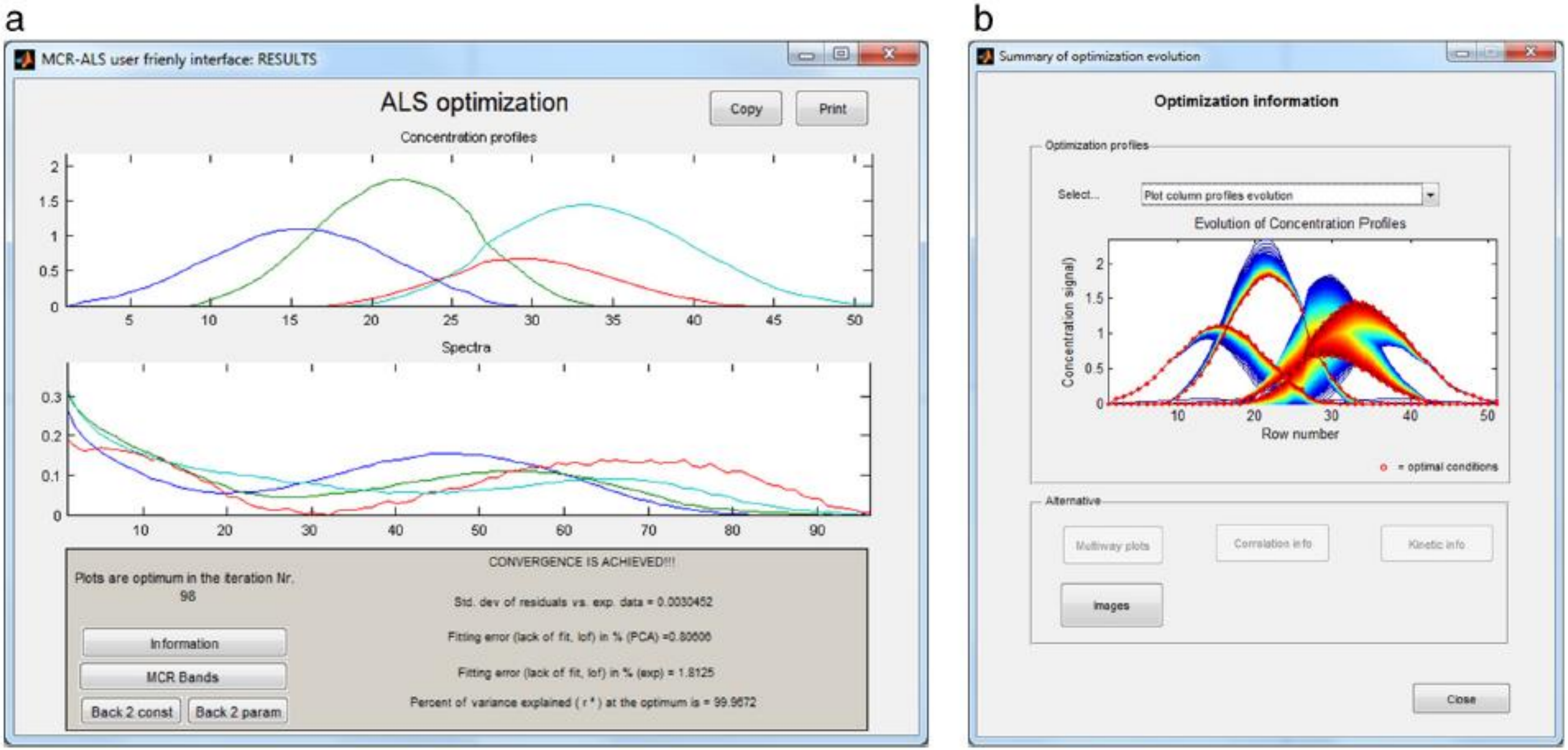
Equality constraints
□ Apply?
Select csel matrix: select a variable from the WS
Constraints are: select...

Advanced constraints
Multiway Correlation Kinetic HM

Reset Continue

http://www.mcrals.info/

3e) Multivariate curve resolution by alternating least-squares (MCR-ALS)



Plots are optimum in the iteration Nr.
98

CONVERGENCE IS ACHIEVED!!!

Std. dev of residuals vs. exp. data = 0.0030452

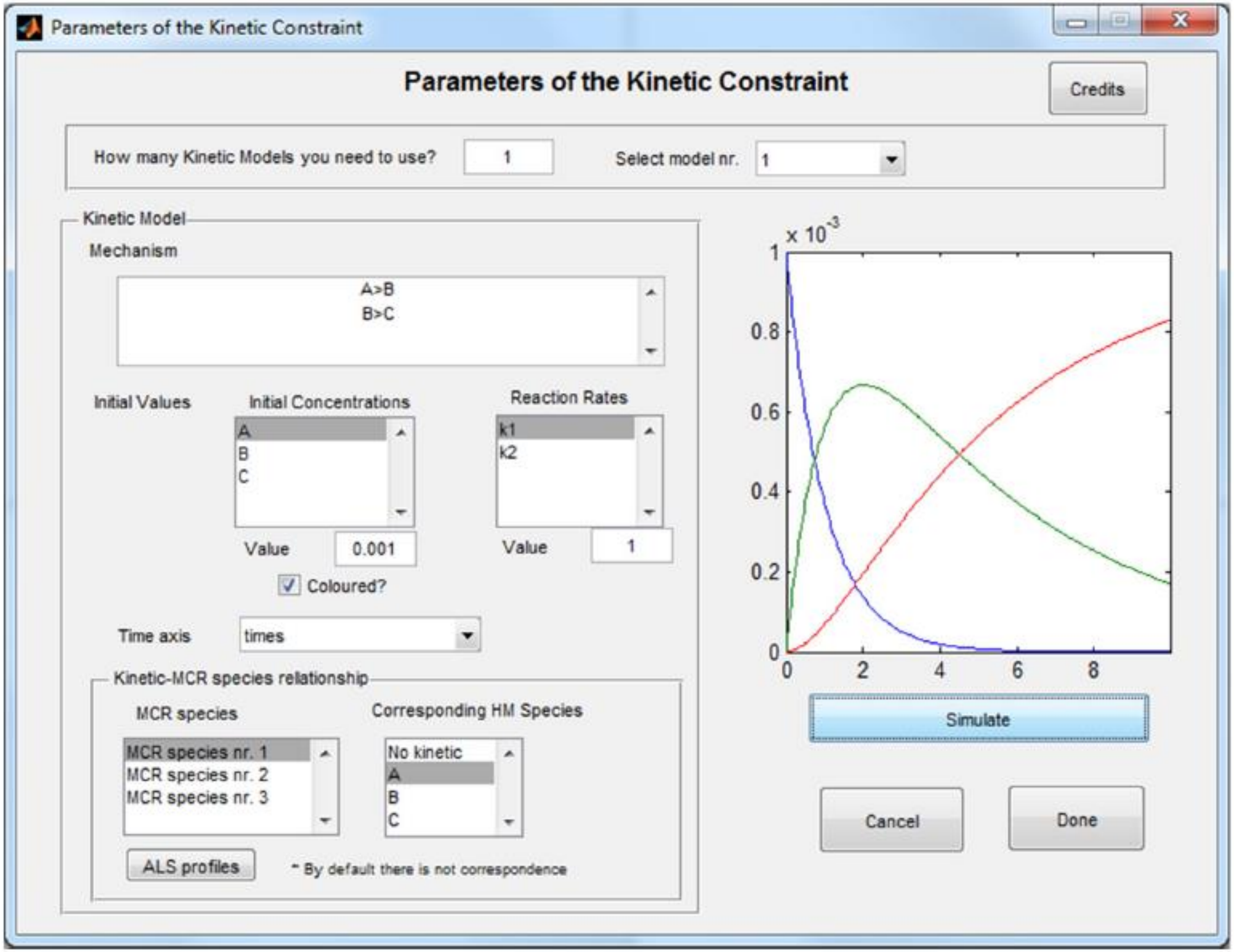
Fitting error (lack of fit, lof) in % (PCA) =0.80606

Fitting error (lack of fit, lof) in % (exp) = 1.8125

Percent of variance explained (r²) at the optimum is = 99.9672

http://www.mcrals.info/

3e) Multivariate curve resolution by alternating least-squares (MCR-ALS)



Parameters of the Kinetic Constraint

How many Kinetic Models you need to use? 1

Select model nr. 1

Kinetic Model

Mechanism

A>B
B>C

Initial Values

Initial Concentrations

A
B
C

Value 0.001

☑ Coloured?

Reaction Rates

k1
k2

Value 1

Time axis: times

Kinetic-MCR species relationship

MCR species

MCR species nr. 1
MCR species nr. 2
MCR species nr. 3

Corresponding HM Species

No kinetic
A
B
C

* By default there is not correspondence

Conclusion: soft-modelling

- Advantages
  - No prior knowledge on the chemical system required
  - Estimation of the number of linearly dependent absorbing species and their approximate evolution from PCA, EFA & ALS
  - Info for the development of a ‘hard’ model
  - ‘Better than nothing’
- Drawbacks
  - No physical model
  - No predictions for other exp. conditions possible
  - Uniqueness of the result is rarely given and difficult to validate

MCR-ALS is a very nice software, I strongly recommend it. http://www.mcrals.info

MCR-ALS becomes a very powerful method when multiple datasets are used.

Table of content

1) Introduction / Framework

2) Instrumentation / Design of Experiment DoE

3) Kinetic soft modelling methods
(FA, EFA, MCR-ALS)

4) Kinetic hard modelling using calibration free method
(Kinetic fitting of spectroscopic data with non-linear regression)

4) Soft modelling versus hard modelling

Factor model

PCA, EFA, MCR-ALS

Soft

Number of species
Approx. conc. & spectra

Lambert-Beer

Kinetic model

$$\upsilon = k \cdot c_{\mathrm{BuOH}} \cdot c_{\mathrm{AA}} \cdot c_{\mathrm{TMG}}$$

$$\frac{dc_{\mathrm{BuOH}}}{dt} = -\upsilon + \frac{f_{in}}{V} \cdot (c_{in,\mathrm{BuOH}} - c_{\mathrm{BuOH}})$$

$$\frac{dc_{\mathrm{AA}}}{dt} = -\upsilon - \frac{f_{in}}{V} \cdot c_{\mathrm{AA}}$$

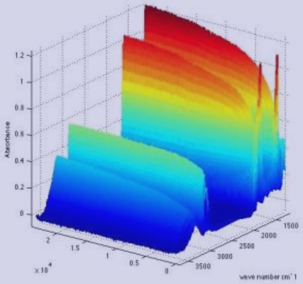
$$\frac{dc_{\mathrm{BuOA}}}{dt} = \upsilon - \frac{f_{in}}{V} \cdot c_{\mathrm{BuOA}}$$

$$\frac{dc_{\mathrm{AH}}}{dt} = \upsilon - \frac{f_{in}}{V} \cdot c_{\mathrm{AH}}$$

$$\frac{dc_{\mathrm{TMG}}}{dt} = -\frac{f_{in}}{V} \cdot c_{\mathrm{TMG}}$$

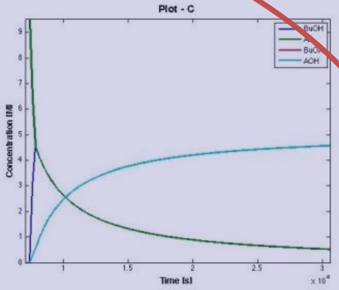
$$\frac{dV}{dt} = f_{in}$$

Reaction Spectra



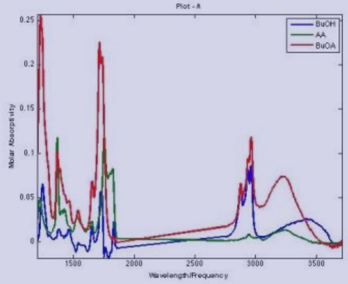
=

Concentrations



×

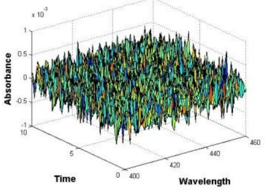
Species Spectra



Hard

Rate constants
Activation energies

+ Residuals

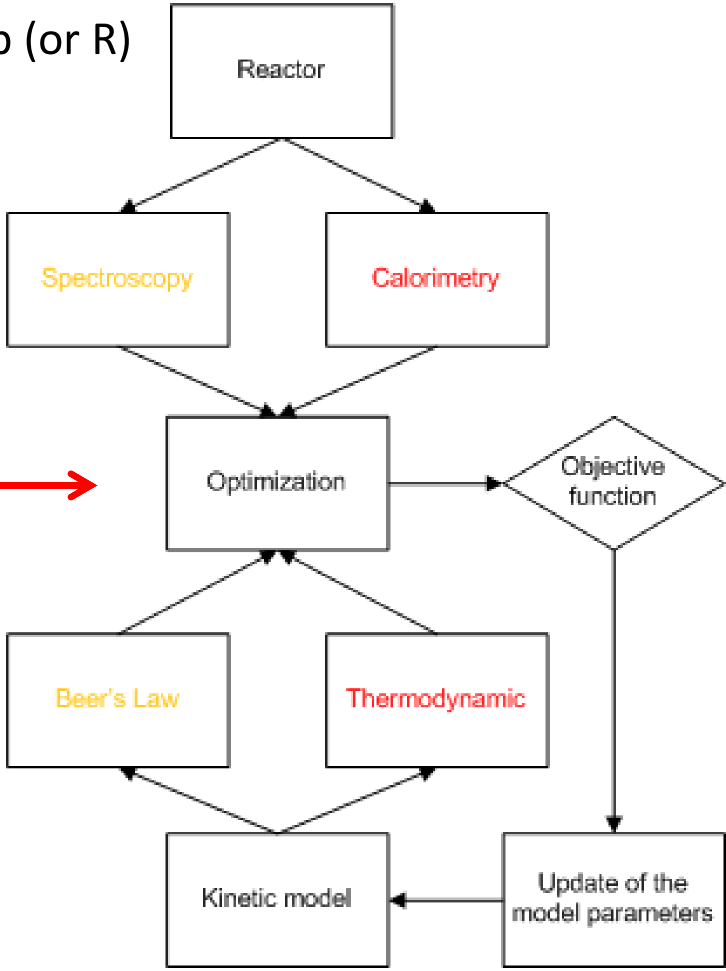


Hard modelling: Use a mathematical model to “explain” the experimental data

4 Kinetic fitting of spectroscopic data with non-linear regression

The objective of this section

- to understand the underlying ideas and principles of the calibration-free modelling (fitting) of spectro-kinetic absorbance data
- to apply these principles to kinetic problems using Matlab (or R)



$\mathbf{Y}_{\mathbf{Measured}} \leftarrow$

4c) $$\min_{\mathbf{k}} \left\| \mathbf{Y}_{\mathbf{Measured}} - \mathbf{C}_{\mathbf{modelled}} \mathbf{A}_{\mathbf{modelled}} \right\|^2 \rightarrow$$

4b) $\mathbf{A}_{\mathbf{modelled}} \leftarrow$ $$\mathbf{A}_{\mathbf{modelled}} = \mathbf{C}^{+}_{\mathbf{modelled}} \mathbf{Y}_{\mathbf{Measured}}$$

4a) $\mathbf{C}_{\mathbf{Modelled}} \leftarrow$ $$\left\{ \begin{aligned} \frac{dC_{Ac_2O}}{dt} &= r(k,t) + \frac{f_{dos}}{V_r(t)} \cdot (C_{dos,Ac_2O} - C_{Ac_2O}) \\ \frac{dC_{H_2O}}{dt} &= r(k,t) - \frac{f_{dos}}{V_r(t)} \cdot C_{H_2O} \\ \frac{dC_{AcOH}}{dt} &= -2 \cdot r(k,t) + \frac{f_{dos}}{V_r(t)} \cdot (C_{dos,AcOH} - C_{AcOH}) \\ \frac{dV_r}{dt} &= f_{dos} \end{aligned} \right.$$

4a) The rate law – modelling concentrations profiles

$$\mathrm{C} = f\,(\text{kinetic model, parameters})$$

rate law

- initial concentrations
- rate constants
- activation energies
- temperature

The rate of a molecular reaction is defined by the derivative of the concentration of the reactants with respect to time normalised by the corresponding stoichiometric coefficient

e.g. $$aA + bB \xrightarrow{k} cC$$

$$\boxed{rate = -\frac{d\left[A\right]_t}{dt \cdot a} = -\frac{d\left[B\right]_t}{dt \cdot b} = \frac{d\left[C\right]_t}{dt \cdot c} = k\left[A\right]_t^a\left[B\right]_t^b}$$

4a) Kinetic model (Batch)

$$A + B \xrightarrow{k_1} C$$

$$2C \xrightarrow{k_2} D$$

Rate laws

$$\boxed{\begin{aligned} rate_1 &= k_1 \cdot \left[A\right]_t \cdot \left[B\right]_t \\ rate_2 &= k_2 \cdot \left[C\right]_t^2 \end{aligned}}$$

$k_1,\ k_2:\ rate\ constants\ [L/mol/s]$

$[A]_t,\ [B]_t,\ [C]_t:\ conc.\ [mol/L]\ of\ species\ at\ time\ t$

Derivatives

$$\boxed{\begin{aligned} \frac{d\left[A\right]_t}{dt} &= \frac{d\left[B\right]_t}{dt} = -rate_1 \\ \frac{d\left[C\right]_t}{dt} &= rate_1 - 2rate_2 \\ \frac{d\left[D\right]_t}{dt} &= rate_2 \end{aligned}}$$

Except for some very simple cases there is no explicit solution for systems of kinetic ODEs ➔ **Numerical integration**

4a) Kinetic model (semi-Batch)

$$A + B \xrightarrow{k_1} C$$

$$2C \xrightarrow{k_2} D$$

When species are dosed, the set of ODEs can be modified accordingly :

Rate laws

$$rate_1 = k_1 \cdot \left[A\right]_t \cdot \left[B\right]_t$$

$$rate_2 = k_2 \cdot \left[C\right]_t^2$$

Derivatives

$$\frac{d\left[A\right]_t}{dt} = -rate_1 + \frac{F_t}{V_t} \cdot \left(\left[A\right]^{dos} - \left[A\right]_t\right)$$

$$\frac{d\left[B\right]_t}{dt} = -rate_1 + \frac{F_t}{V_t} \cdot \left(\left[B\right]^{dos} - \left[B\right]_t\right)$$

$$\frac{d\left[C\right]_t}{dt} = rate_1 - 2 \cdot rate_2 + \frac{F_t}{V_t} \cdot \left(\left[C\right]^{dos} - \left[C\right]_t\right)$$

$$\frac{d\left[D\right]_t}{dt} = rate_2 + \frac{F_t}{V_t} \cdot \left(\left[D\right]^{dos} - \left[D\right]_t\right)$$

$$\frac{dV_t}{dt} = F_t$$

F_t : *flow rate* $[mL/s]$ *at time t*

V_t : *volume* $[mL]$ *at time t*

$[A]^{dos}$: *conc.* $[mol/L]$ *of dosed species*

dosed material: $+\frac{F_t}{V_t} \cdot [A]^{dos}$

dilution: $-\frac{F_t}{V_t} \cdot [A]_t$

Dosing requires some modifications to the ODEs for all species and the inclusion of an ODE for the change of volume due to the flow-rate

4a) Numerical integration of the rate law

- Crude approach : Euler’s method (truncated Taylor series)

$$c_i(t+\Delta t) \approx c_i(t) + \left(\frac{dc_i}{dt}\right)_t \cdot \Delta t$$

Applied to our specific example *without dosing*

$$[C]_{t+\Delta t} \approx [C]_t + \left(\frac{d[C]_t}{dt}\right)_t \cdot \Delta t$$

$$\approx [C]_t + \left(k_1 \cdot [A]_t \cdot [B]_t - 2k_2 \cdot [C]_t^2\right)\Delta t$$

$$A + B \xrightarrow{k_1} C$$

$$2C \xrightarrow{k_2} D$$

- Much more sophisticated integration methods exist, e.g. Matlab’s ‘`ode45`’, a 4th order Runge-Kutta with ***automatic stepsize control***

4a) Numerical integration of the rate law

- In stepsize controlled ODE solvers, the stepsize is adjusted at each step to meet the user-specified accuracy

- The accuracy is measured with absolute (Matlab: **`AbsTol`**) and relative (Matlab: **`RelTol`**) tolerance's values.

- For some kinetic models, the concentration profiles change on dramatically different scales (stiff problem) and a **stiff ODE solver** (eg. Matlab's **`‘ode15s’`**) is required

4a) Solving ODE's with Matlab: Batch example

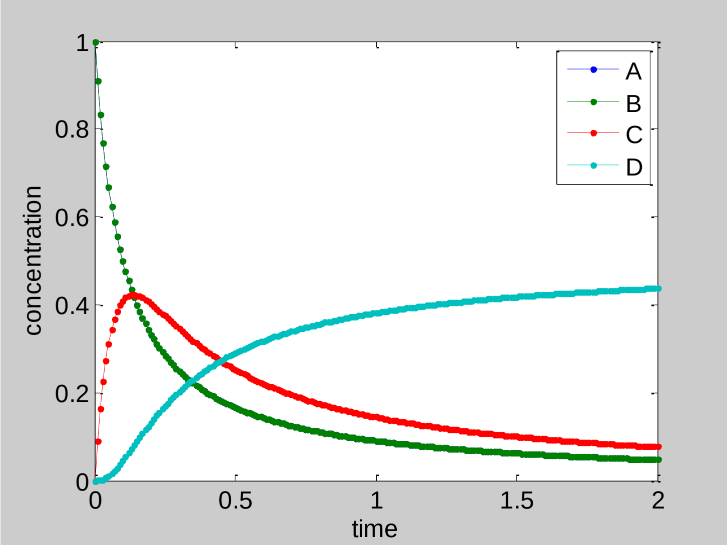
```
% Mechanism : A+B>C, 2C>D (batch)
t          = [0:0.01:2]';        % time vector (column)
c0         = [1 1 0 0];          % initial concentrations (row)
k          = [10; 5];            % rate constants (column)

% call ODE solver for numerical integration of dC/dt
odeoptions = odeset('RelTol',1e-10,'AbsTol',1e-12);
[tout,C] = ode45('ode_ApBtoC_2CtoD_batch',t,c0',odeoptions,k);
```

$$A + B \xrightarrow{k_1} C$$

$$2C \xrightarrow{k_2} D$$

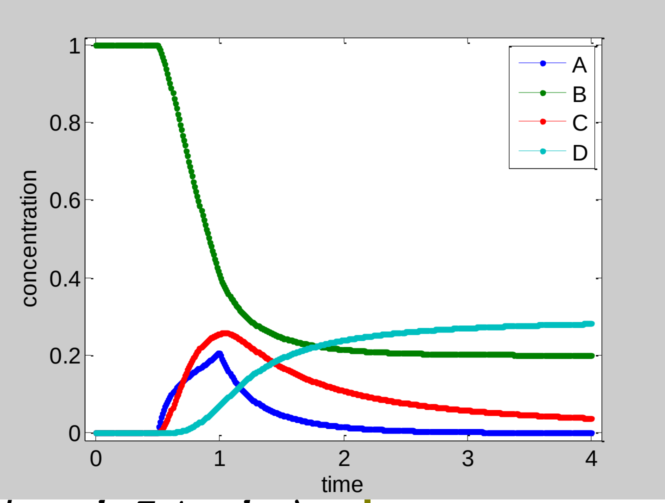
batch conditions

```
function cdot = ode_ApBtoC_2CtoD_batch(ti,ci,flag,k)
% Mechanism : A+B>C, 2C>D (batch)

% rate law
rates(1)     = k(1)*ci(1)*ci(2);
rates(2)     = k(2)*ci(3)*ci(3);

% System of ODEs
cdot(1,1)    = -rates(1);                % dA/dt
cdot(2,1)    = -rates(1);                % dB/dt
cdot(3,1)    =  rates(1) - 2*rates(2);   % dC/dt
cdot(4,1)    =  rates(2);                % dD/dt
```



4a) Solving ODE's with Matlab: semi-batch example

```
% Mechanism : A+B>C, 2C>D (semi batch)
t      = [0:0.01:4]'; % time vector (column)
c0     = [0 1 0 0];    % initial concentrations (row)
k      = [10;5];       % rate constants (column)
V0     =  0.020;       % initial volume [L]
tdos   = [0.5 1];      % start and stop times of dosing (row)
cdos   = [3 0 0 0];    % concentrations of dosed species (row)
frate =  0.01;         % flow-rate [L/(unit of time)]
F = flow(t,tdos,frate);   % generate vector F of flow-rates

% call ODE solver for numerical integration of dC/t and dV/t
odeoptions = odeset('RelTol',1e-10,'AbsTol',1e-12);
[tout,CV]  = ode45('ode_ApBtoC_2CtoD_semi',t,[c0 V0]',odeoptions,k,F,t,cdos);
% extract volume and concentration profiles from columns of matrix CV
C = CV(:,1:end-1); V = CV(:,end);
```



```
function cdot = ode_ApBtoC_2CtoD_semi(ti,civi,flag,k,F,t,cdos)
% Mechanism : A+B>C, 2C>D (semi batch)
ci          = civi(1:end-1);             % concentration at time ti
Vi          = civi(end);                 % volume at time ti
fi          = F(find(t<=ti,1,'last')); % flow-rate at time ti
% rate law
rates(1)   = k(1)*ci(1)*ci(2);
rates(2)   = k(2)*ci(3)*ci(3);
% System of ODEs
cdot(1,1)  = -rates(1)                + (fi/Vi)*(cdos(1)-ci(1)); % dA/dt
cdot(2,1)  = -rates(1)                + (fi/Vi)*(cdos(2)-ci(2)); % dB/dt
cdot(3,1)  =  rates(1) - 2*rates(2) + (fi/Vi)*(cdos(3)-ci(3)); % dC/dt
cdot(4,1)  =  rates(2)                + (fi/Vi)*(cdos(4)-ci(4)); % dD/dt
cdot(5,1)  =  fi;                                                 % dV/dt
```

$$A + B \xrightarrow{k_1} C$$

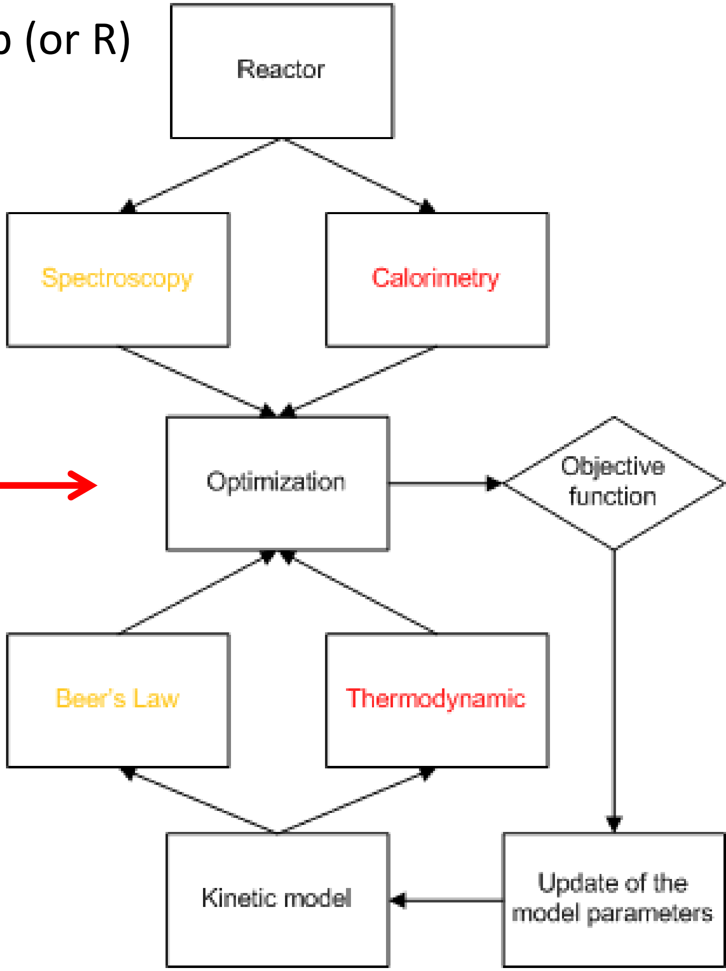
$$2C \xrightarrow{k_2} D$$

semi-batch
(*A* dosed)

4 Kinetic fitting of spectroscopic data with non-linear regression

The objective of this section

- to understand the underlying ideas and principles of the calibration-free modelling (fitting) of spectro-kinetic absorbance data
- to apply these principles to kinetic problems using Matlab (or R)



$\mathbf{Y}_{\text{Measured}} \leftarrow$

4c) $$\min_{k} \left\| \mathbf{Y}_{\text{Measured}} - \mathbf{C}_{\text{modelled}} \mathbf{A}_{\text{modelled}} \right\|^2 \rightarrow$$

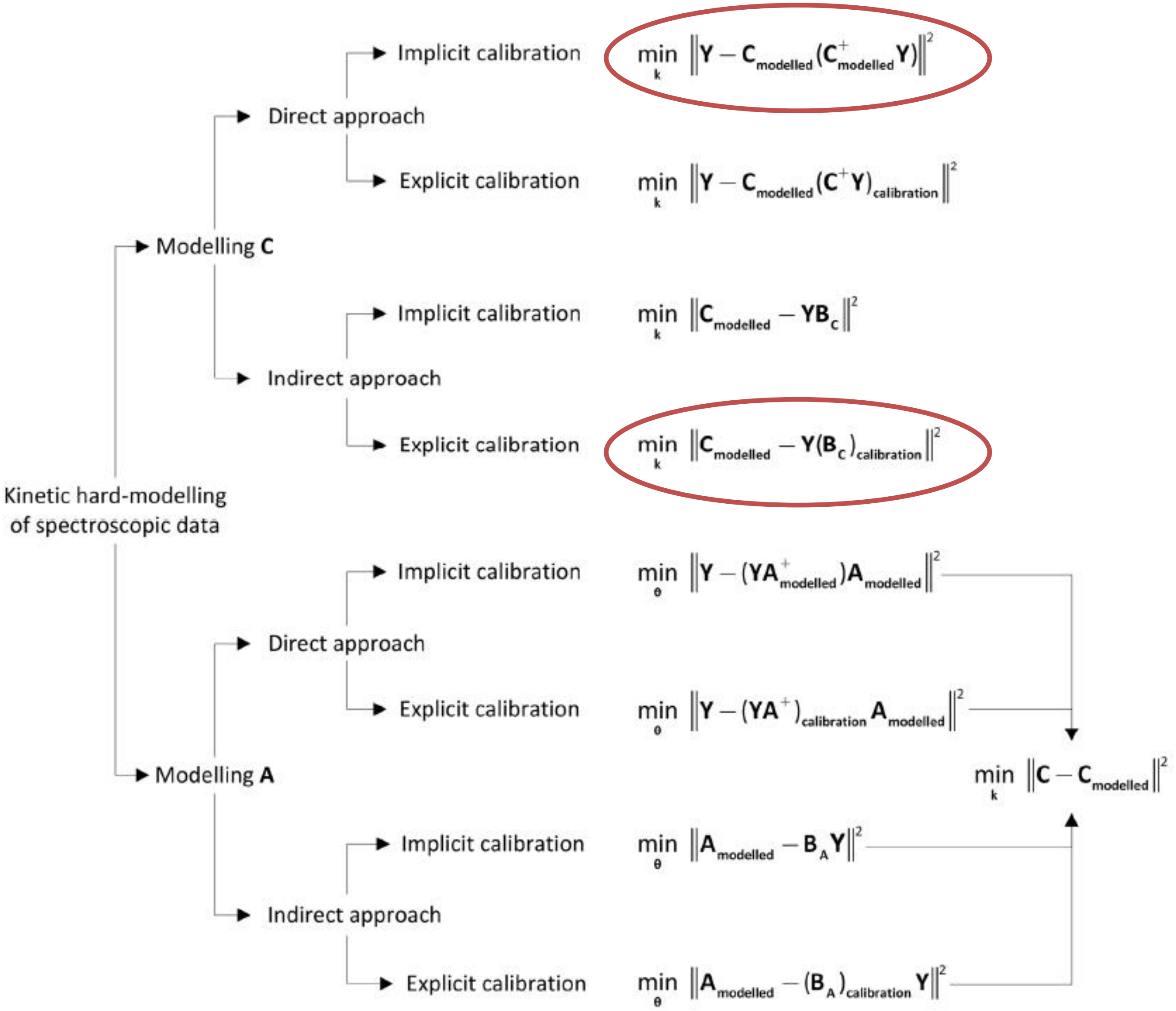
4b) $\mathbf{A}_{\text{modelled}} \leftarrow$ $\mathbf{A}_{\text{modelled}} = \mathbf{C}^{+}_{\text{modelled}} \mathbf{Y}_{\text{Measured}}$

4a) $\mathbf{C}_{\text{Modelled}} \leftarrow$

$$\begin{cases} \frac{dC_{Ac_2O}}{dt} &= r(k,t) + \frac{f_{dos}}{V_r(t)} \cdot (C_{dos,Ac_2O} - C_{Ac_2O}) \\ \frac{dC_{H_2O}}{dt} &= r(k,t) - \frac{f_{dos}}{V_r(t)} \cdot C_{H_2O} \\ \frac{dC_{AcOH}}{dt} &= -2 \cdot r(k,t) + \frac{f_{dos}}{V_r(t)} \cdot (C_{dos,AcOH} - C_{AcOH}) \\ \frac{dV_r}{dt} &= f_{dos} \end{cases}$$

4b) Selection of the objective function?

Decision tree for the selection of an appropriate method for the kinetic hard-modelling of spectroscopic data



Kinetic hard-modelling of spectroscopic data

Modelling **C**

Direct approach

Implicit calibration $\min_{k} \left\| \mathbf{Y} - \mathbf{C}_{modelled}(\mathbf{C}^{+}_{modelled}\mathbf{Y}) \right\|^2$

Explicit calibration $\min_{k} \left\| \mathbf{Y} - \mathbf{C}_{modelled}(\mathbf{C}^{+}\mathbf{Y})_{calibration} \right\|^2$

Indirect approach

Implicit calibration $\min_{k} \left\| \mathbf{C}_{modelled} - \mathbf{Y}\mathbf{B}_{C} \right\|^2$

Explicit calibration $\min_{k} \left\| \mathbf{C}_{modelled} - \mathbf{Y}(\mathbf{B}_{C})_{calibration} \right\|^2$

Modelling **A**

Direct approach

Implicit calibration $\min_{\theta} \left\| \mathbf{Y} - (\mathbf{Y}\mathbf{A}^{+}_{modelled})\mathbf{A}_{modelled} \right\|^2$

Explicit calibration $\min_{\theta} \left\| \mathbf{Y} - (\mathbf{Y}\mathbf{A}^{+})_{calibration}\mathbf{A}_{modelled} \right\|^2$

Indirect approach

Implicit calibration $\min_{\theta} \left\| \mathbf{A}_{modelled} - \mathbf{B}_{A}\mathbf{Y} \right\|^2$

Explicit calibration $\min_{\theta} \left\| \mathbf{A}_{modelled} - (\mathbf{B}_{A})_{calibration}\mathbf{Y} \right\|^2$

$\min_{k} \left\| \mathbf{C} - \mathbf{C}_{modelled} \right\|^2$

4b) the objective function

Target: Find the least-squares minimum as a function of the rate constant(s), the non-linear parameters

for 'concentration measurements', $\mathbf{C}$

$$ssq(\mathbf{k}) = \sum_{i=1}^{n_t} \sum_{s=1}^{n_s} r_{i,s}^2(\mathbf{k}), \qquad \mathbf{R}(\mathbf{k}) = \mathbf{C} - \mathbf{C}_{\mathbf{calc}}(\mathbf{k})$$

for spectral absorbance measurements, $\mathbf{Y}$

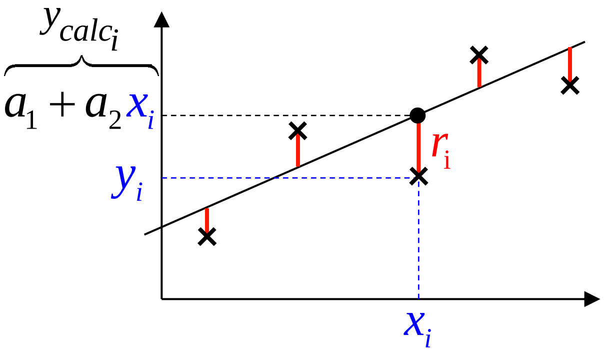
$$ssq(\mathbf{k}) = \sum_{i=1}^{n_t} \sum_{j=1}^{n_\lambda} r_{i,j}^2(\mathbf{k}), \qquad \mathbf{R}(\mathbf{k}) = \mathbf{Y} - \mathbf{Y}_{\mathbf{calc}}(\mathbf{k}) = \mathbf{Y} - \mathbf{C}_{\mathbf{calc}}(\mathbf{k}) \cdot \mathbf{A}_{\mathbf{calc}}(\mathbf{k})$$

$$\partial ssq \big/ \partial \mathbf{k} = 0$$

No explicit solution !!!
→ problem to be solved iteratively

4b) Data fitting – a quick review of basic principles

- The most widely used data fitting is the linear regression of a data vector $\mathbf{y}$ to a straight line (1st order polynomial)



$$y_i = a_1 + a_2 x_i + r_i$$

$$i = 1 \ldots m,\ m > 2$$

- As with all other regression methods the task is to minimise the least-squares sum *ssq* of all residuals r_i between the data y_i and the underlying model $a_1 + a_2 x_i$ by optimising the linear parameters defining the model, here slope a_2 and intercept a_1

$$\min_{a_1,\, a_2} |ssq| \qquad \text{where} \qquad ssq(a_1, a_2) = \sum_{i=1}^{m} r_i^2 = \sum_{i=1}^{m} \left(y_i - \underbrace{(a_1 + a_2 x_i)}_{y_{calc_i}} \right)^2$$

4b) Data fitting – a quick review of basic principles

- In order to find the “best” parameters a_1 & a_2, i.e. that lead to a minimal ssq the following two derivatives must be zero:

$$\partial ssq / \partial a_1 = \partial ssq / \partial a_2 = 0$$

- For linear parameters, such as slope and intercept there is a non-iterative solution, i.e. for m data points, a_1 and a_2 can be calculated explicitly:

$$a_1 = \frac{\Sigma x_i^2 \Sigma y_i - \Sigma x_i \Sigma x_i y_i}{m \Sigma x_i^2 - (\Sigma x_i)^2}$$

$$a_2 = \frac{-\Sigma x_i \Sigma y_i + m \Sigma x_i y_i}{m \Sigma x_i^2 - (\Sigma x_i)^2}$$

- But linear regression is much more than just straight line fitting. It also includes the fitting to polynomials of any other order or any other function that depends on linear parameters only.

4b) Data fitting – a quick review of basic principles

- Linear relationships can be written much simpler in an elegant vector or matrix notation. The straight line regression problem is then denoted by:

$$\begin{aligned} y_1 &= a_1 + a_2 x_1 + r_1 \\ y_2 &= a_1 + a_2 x_2 + r_2 \\ &\vdots \\ y_i &= a_1 + a_2 x_i + r_i \\ &\vdots \\ y_m &= a_1 + a_2 x_m + r_m \end{aligned} \longrightarrow \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} + \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_i \\ \vdots \\ r_m \end{bmatrix}$$

$$\mathbf{y} = \mathbf{F}(\mathbf{x}) \cdot \mathbf{a} + \mathbf{r}$$

$\mathbf{F}(\mathbf{x})$ is the design matrix

4b) Data fitting – a quick review of basic principles

- And the generalisation for any polynomial

$$y_i = a_1 + a_2 x_i + a_3 x_i^2 + \ldots + a_j x_i^{j-1} + \ldots + a_{n_a} x_i^{n_a - 1} = \sum_{j=1}^{n_a} a_j x_i^{j-1}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_i \\ \vdots \\ y_m \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^{j-1} & \cdots & x_1^{n_a-1} \\ 1 & x_2 & x_2^2 & \cdots & x_2^{j-1} & \cdots & x_2^{n_a-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots \\ 1 & x_i & x_i^2 & \cdots & x_i^{j-1} & \cdots & x_i^{n_a-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & x_m^2 & \cdots & x_m^{j-1} & \cdots & x_m^{n_a-1} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_j \\ \vdots \\ a_{n_a} \end{bmatrix} + \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_i \\ \vdots \\ r_m \end{bmatrix}$$

$$\mathbf{y} = \mathbf{F}(\mathbf{x})\,\mathbf{a} + \mathbf{r}$$

4b) Beer's law: Linear relationship between spectral absorbance and species concentrations

Calibration with pure spectra and fitting ***C****:*

- Analogously, if the pure species spectra $\mathbf{A}$ are known the corresponding concentration profiles $\mathbf{C}$ can also be determined by **multivariate linear regression**:

$$\mathbf{Y} = \mathbf{C} \cdot \mathbf{A} + \mathbf{R} \qquad \text{such that } ssq = \sum_{i=1}^{n_t} \sum_{j=1}^{n_\lambda} r_{i,j}^2 \text{ is minimal}$$

$$\mathbf{C} = \mathbf{Y}\mathbf{A}^{\mathrm{t}}(\mathbf{A}\mathbf{A}^{\mathrm{t}})^{-1}$$

$\mathbf{A}^{\mathrm{t}}(\mathbf{A}\mathbf{A}^{\mathrm{t}})^{-1}$ is called the right pseudoinverse, $\mathbf{A}^{+}$, of $\mathbf{A}$

Calibration free and fitting ***Y****:*

- For a given multi wavelength kinetic measurement, $\mathbf{Y}$, if the concentration profiles $\mathbf{C}$ are known the corresponding pure species spectra $\mathbf{A}$ can be determined by **multivariate linear regression**:

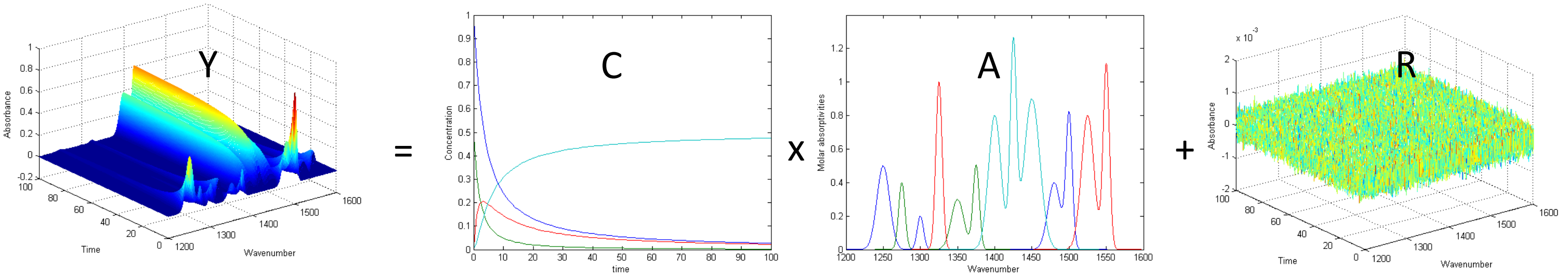
$$\mathbf{Y} = \mathbf{C} \cdot \mathbf{A} + \mathbf{R} \qquad \text{such that } ssq = \sum_{i=1}^{n_t} \sum_{j=1}^{n_\lambda} r_{i,j}^2 \text{ is minimal}$$

$$\mathbf{A} = (\mathbf{C}^{\mathrm{t}}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{t}}\mathbf{Y}$$

$(\mathbf{C}^{\mathrm{t}}\mathbf{C})^{-1}\mathbf{C}^{\mathrm{t}}$ is called the left pseudoinverse, $\mathbf{C}^{+}$, of $\mathbf{C}$

4b) Beer's law: Linear relationship between spectral absorbance and species concentrations

Beer's law in matrix notation



With calibration

$$\mathbf{A}_{\mathit{Calibration}} = \text{function } (\boldsymbol{\theta})$$

$$\mathbf{C} = (\mathbf{YA}^{+}_{\mathbf{calibration}})$$

Calibration free

$$\mathbf{C}_{\mathbf{modelled}} = \text{function } (\mathbf{k})$$

$$\mathbf{A} = \mathbf{C}^{+}_{\mathbf{modelled}}\mathbf{Y} \qquad [1]$$

Objective function:

$$\min_{\mathbf{k}} \left\|\mathbf{C}_{\mathbf{Measured}} - \mathbf{C}_{\mathbf{modelled}}\right\|^2 \qquad \min_{\mathbf{k}} \left\|\mathbf{Y}_{\mathbf{Measured}} - \mathbf{C}_{\mathbf{modelled}}\mathbf{A}\right\|^2$$

[1] $\mathbf{C}^{+} = (\mathbf{C}^{T}\mathbf{C})^{-1}\mathbf{C}^{T}$

4 Kinetic fitting of spectroscopic data with non-linear regression

The objective of this section

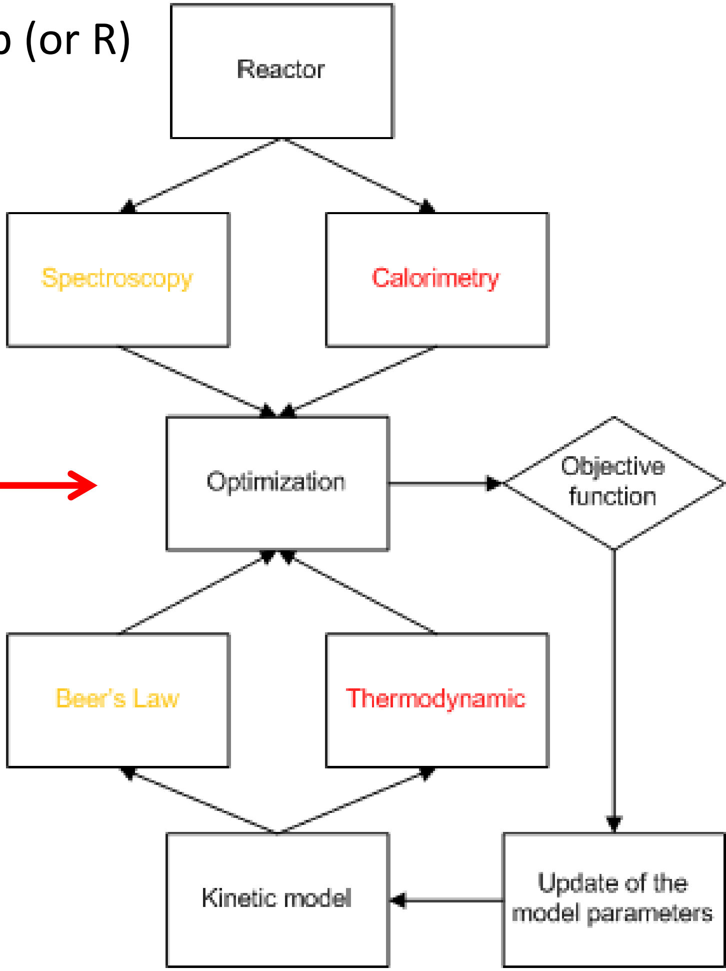
- to understand the underlying ideas and principles of the calibration-free modelling (fitting) of spectro-kinetic absorbance data
- to apply these principles to kinetic problems using Matlab (or R)

$Y_{Measured}$ ←

4c) $$\min_{k} \left\| Y_{Measured} - C_{modelled} A_{modelled} \right\|^2$$ →

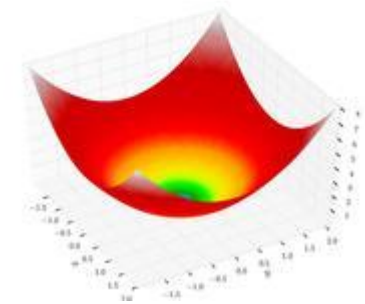
4b) $A_{modelled}$ ← $$A_{modelled} = C^{+}_{modelled} Y_{Measured}$$

4a) $C_{Modelled}$ ← $$\begin{cases} \frac{dC_{Ac_2O}}{dt} & = r(k,t) + \frac{f_{dos}}{V_r(t)} \cdot (C_{dos,Ac_2O} - C_{Ac_2O}) \\ \frac{dC_{H_2O}}{dt} & = r(k,t) - \frac{f_{dos}}{V_r(t)} \cdot C_{H_2O} \\ \frac{dC_{AcOH}}{dt} & = -2 \cdot r(k,t) + \frac{f_{dos}}{V_r(t)} \cdot (C_{dos,AcOH} - C_{AcOH}) \\ \frac{dV_r}{dt} & = f_{dos} \end{cases}$$

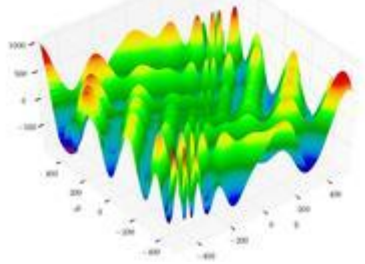


4c) Fitting of kinetics models to concentration profiles or absorbance data

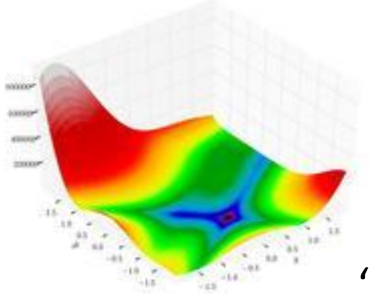
Sphere:



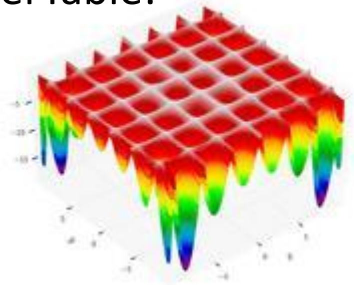
Eggholder:



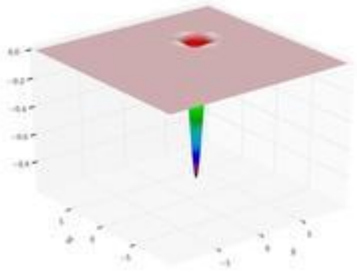
Goldstein-Price:



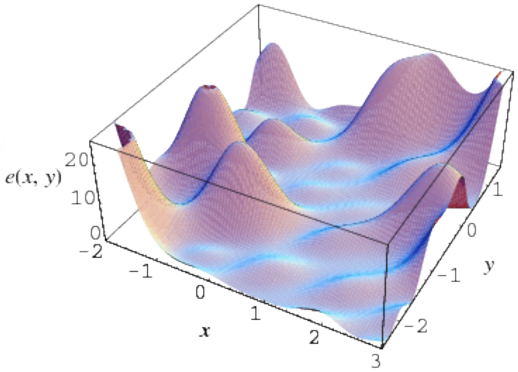
HölderTable:



Easom:



“Real case”:



Global optimizer (Heuristic algoritm):

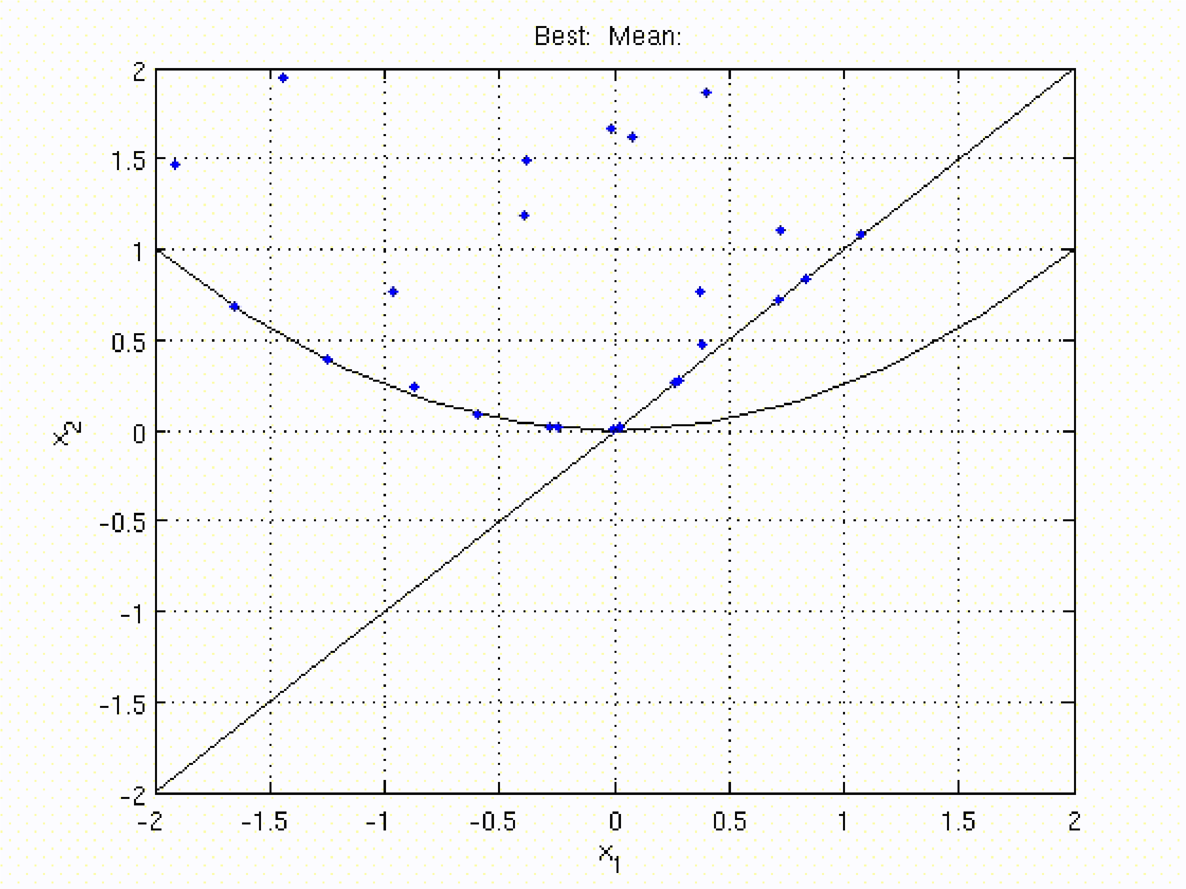
- Evolutionary algorithm
- Genetic algorithms (GA)
- Simulated annealing (SA)
- Tabu search
- Memetic algorithm
- Particle swarm optimization (PSO)

Iterative (gradient based methods)

- Sequential quadratic programming (SQP, Hessian ev.)
- Quasi-Newton methods
- Gradient descent
- Simplex
- Simultaneous perturbation stochastic approximation (SPSA)

4c) Fitting of kinetics models to concentration profiles or absorbance data

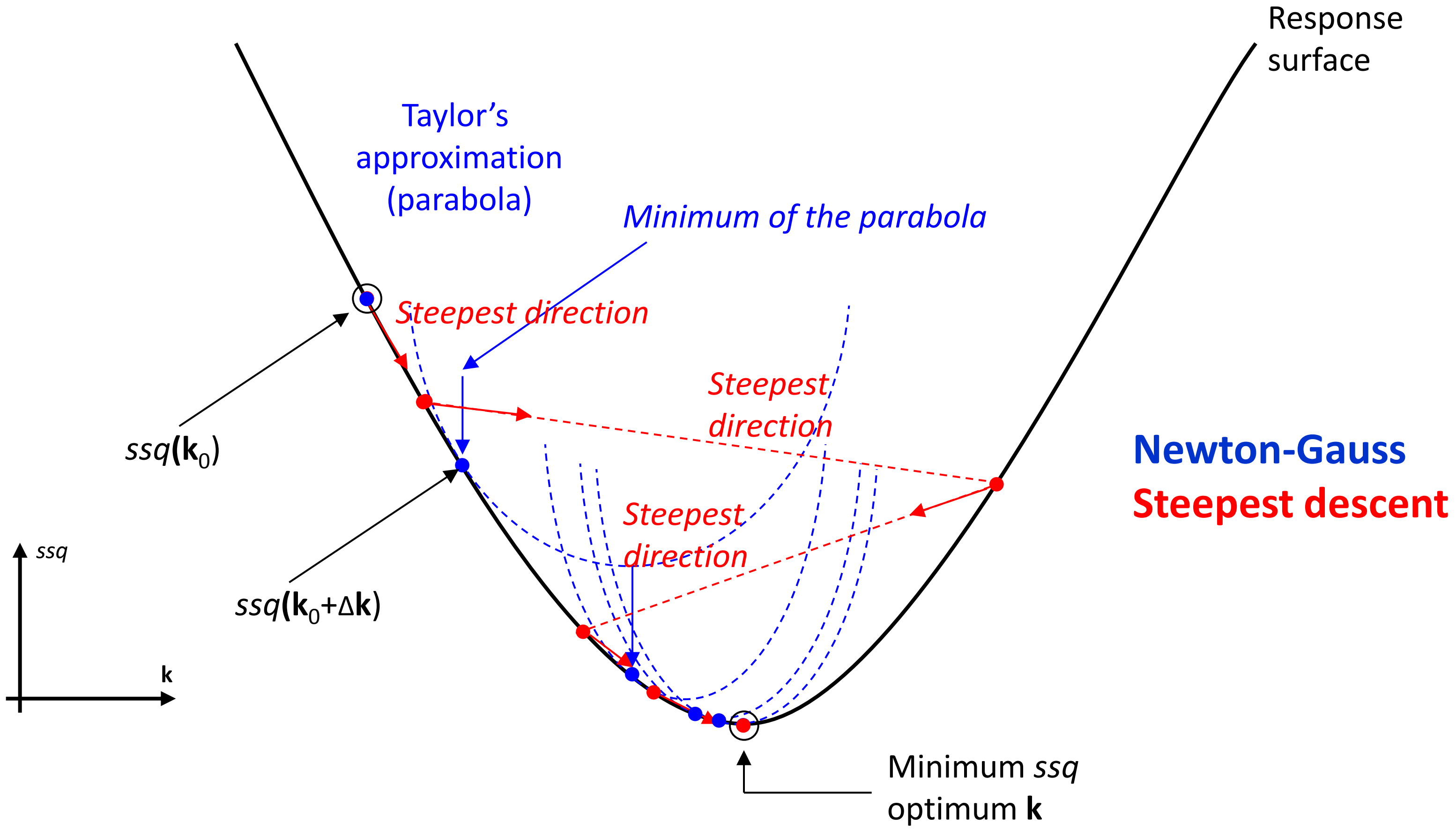
Global optimizer Ex. Particle Swarm Optimisation (PSO)



https://www.youtube.com/watch?v=3CR5y8qZf0Y

4c) Fitting of kinetics models to concentration profiles or absorbance data

Iterative (gradient based methods)



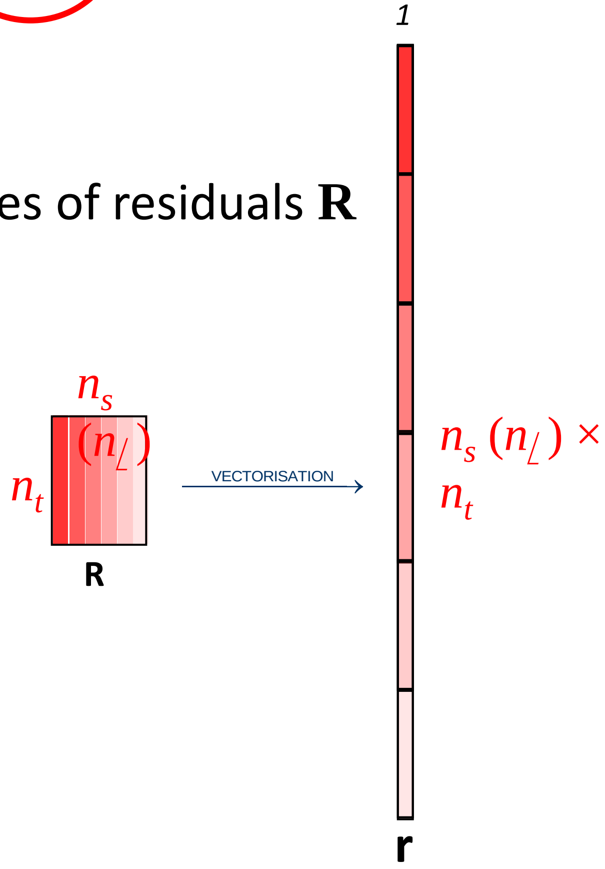
4c) Fitting of kinetics models to concentration profiles or absorbance data

- It is possible to circumvent the calculation of

$$ssq(\mathbf{k}+\Delta\mathbf{k}) \approx ssq(\mathbf{k}) + \left(\frac{\partial ssq}{\partial \mathbf{k}}\right)\cdot\Delta\mathbf{k} + \frac{1}{2}\overset{\text{Hessian}}{\boxed{\left(\frac{\partial^2 ssq}{\partial\mathbf{k}\partial\mathbf{k}}\right)}}\cdot\Delta\mathbf{k}^2$$

and to develop the Taylor series for the residuals $\mathbf{R}$

- To do so it is convenient to first vectorise the matrices of residuals $\mathbf{R} = \mathbf{C}\text{-}\mathbf{C}_{\mathbf{calc}}$ (fitting concentration data $\mathbf{C}$)
  $\mathbf{R}=\mathbf{Y}\text{-}\mathbf{Y}_{\mathbf{calc}}=\mathbf{Y}\text{-}\mathbf{C}_{\mathbf{calc}}\mathbf{A}_{\mathbf{calc}}$ (fitting spectral data $\mathbf{Y}$)



VECTORISATION :

$\mathbf{R}$ is unfolded into a ´long´ column vector $\mathbf{r}$

<u>in Matlab:</u> `r=R(:)`

4c) Fitting of kinetics models to concentration profiles or absorbance data

- **Computationally easier and (almost) equivalent:** The residuals are approximated by a Taylor series expansion truncated after the first derivative

Jacobian J

$$\mathbf{r}(\mathbf{k}+\Delta\mathbf{k}) = \mathbf{r}(\mathbf{k}) + \frac{\partial \mathbf{r}(\mathbf{k})}{\partial \mathbf{k}} \cdot \Delta\mathbf{k}$$

| rearrange for $\mathbf{r}(\mathbf{k})$

$$\mathbf{r}(\mathbf{k}) = -\mathbf{J}\cdot\Delta\mathbf{k} + \mathbf{r}(\mathbf{k}+\Delta\mathbf{k})$$

same structure as before in

$\mathbf{y} = \mathbf{C}\times\mathbf{a} + \mathbf{r}$, with $\mathbf{a} = \mathbf{C}^{+}\mathbf{y} = (\mathbf{C}^{t}\mathbf{C})^{-1}\mathbf{C}^{t}\mathbf{y}$

| *thus: linear regression to minimize* $\mathbf{r}(\mathbf{k}+\Delta\mathbf{k})$ *by optimising* $\Delta\mathbf{k}$ for a given $\mathbf{J}$

$$\boxed{\Delta\mathbf{k} = -(\mathbf{J}^{t}\mathbf{J})^{-1}\mathbf{J}^{t}\cdot\mathbf{r}(\mathbf{k}) = -\mathbf{J}^{+}\cdot\mathbf{r}(\mathbf{k})}$$

The **SHIFT VECTOR** ⊗$\mathbf{k}$ is added to $\mathbf{k}$ for the next iteration

approx. Hessian $\mathbf{H}$

```
Matlab: Δk = -J \ r
```

The Newton-Gauss algorithm requires the calculation of the Jacobian

$$\mathbf{J} = \frac{\partial \mathbf{r}(\mathbf{k})}{\partial \mathbf{k}}$$

4c) Convergence criterion in Newton-Gauss algorithm

The NG algorithm **CONVERGES** if the Taylor series expansion is a good approximation for the new residuals $\mathbf{r}(\mathbf{k}+\Delta\mathbf{k})$, and the shift vector

$$\boxed{\Delta\mathbf{k} = -(\mathbf{J}^{\mathrm{t}}\mathbf{J})^{-1}\mathbf{J}^{\mathrm{t}}\cdot\mathbf{r}(\mathbf{k}) = -\mathbf{J}^{+}\cdot\mathbf{r}(\mathbf{k})}$$

leads to a better (smaller) least-squares sum $ssq(\mathbf{k}+\Delta\mathbf{k})$ than $ssq(\mathbf{k})$ of the previous iteration or with the initial guess of $\mathbf{k}$.

Convergence criterion: “small” relative change of ssq (e.g. μ = 10^{-4})

$$\frac{ssq(\mathbf{k})-ssq(\mathbf{k}+\Delta\mathbf{k})}{ssq(\mathbf{k})} > \mu$$ positive, → convergence

$$\frac{ssq(\mathbf{k})-ssq(\mathbf{k}+\Delta\mathbf{k})}{ssq(\mathbf{k})} < -\mu$$ negative, → divergence

$$abs\left(\frac{ssq(\mathbf{k})-ssq(\mathbf{k}+\Delta\mathbf{k})}{ssq(\mathbf{k})}\right) \leq \mu$$ approx. equal → minimum reached

4c) Convergence criterion in Newton-Gauss algorithm

PROBLEM : The NG algorithm **DIVERGES** if the Taylor series expansion is not a good approximation for the residuals function (e.g. with poor initial guesses of $\mathbf{k}$)

SOLUTION : Do not use a Taylor series expansion but move in the direction of steepest descent

$$\Delta\mathbf{k} = -\mathbf{H}^{-1} \cdot \mathbf{J}^{t} \cdot \mathbf{r}(\mathbf{k})$$

Inverse Hessian method
(Newton-Gauss)

Is there a way to switch progressively from one method to the other ?

$$\boxed{\Delta\mathbf{k} = -(\mathbf{H} + mp \cdot \mathbf{I})^{-1} \cdot \mathbf{J}^{t} \cdot \mathbf{r}(\mathbf{k})}$$

Levenberg-Marquardt
modification

The Marquardt parameter (mp) is a scalar added to the diagonal elements of $\mathbf{H}$ to decrease its influence on $\otimes\mathbf{k}$ and shorten the magnitude of $\otimes\mathbf{k}$

$$\Delta\mathbf{k} = -\mathbf{J}^{t} \cdot \mathbf{r}(\mathbf{k})$$

steepest descent

4c) The Hessian (Evaluation of the error statistics)

- With $\mathbf{H} = \mathbf{J}^{t}{,}\mathbf{J}$ the shift vector Δ$\mathbf{k}$ can be written as:

$$\Delta\mathbf{k} = -\mathbf{H}^{-1}\mathbf{J}^{t}\mathbf{r}\left(\mathbf{k}\right)$$

The Hessian $\mathbf{H}$ is a square matrix ($n_k \times n_k$). It’s inverse is an estimate for the variance/covariance matrix for $\mathbf{k}$!

The diagonal element(s) of the inverted Hessian allow the calculation of the **standard error(s)** $\sigma_{\mathbf{k}}$ for the rate constant(s) $\mathbf{k}$:

$$\sigma_{\mathbf{k}} = \sigma_{\mathbf{r}} \cdot \sqrt{diag\left(\mathbf{H}^{-1}\right)} \qquad with \quad \sigma_{\mathbf{r}} = \sqrt{\frac{ssq}{df}} \approx \sigma_{\mathbf{y}}$$

$\sigma_{\mathbf{r}}$: standard deviation of the residuals $\mathbf{r}$ ($\mathbf{R}$)

$\sigma_{\mathbf{y}}$: ´true´ standard deviation of the measurement $\mathbf{Y}$ (or $\mathbf{C}$)

df: degree of freedom, $df = n_t n_s - n_k$ ($\mathbf{C}$ fitted), or $df = n_t n_\lambda - n_k - n_s n_\lambda$($\mathbf{Y}$ fitted)

4 Kinetic fitting of spectroscopic data with non-linear regression

The objective of this section

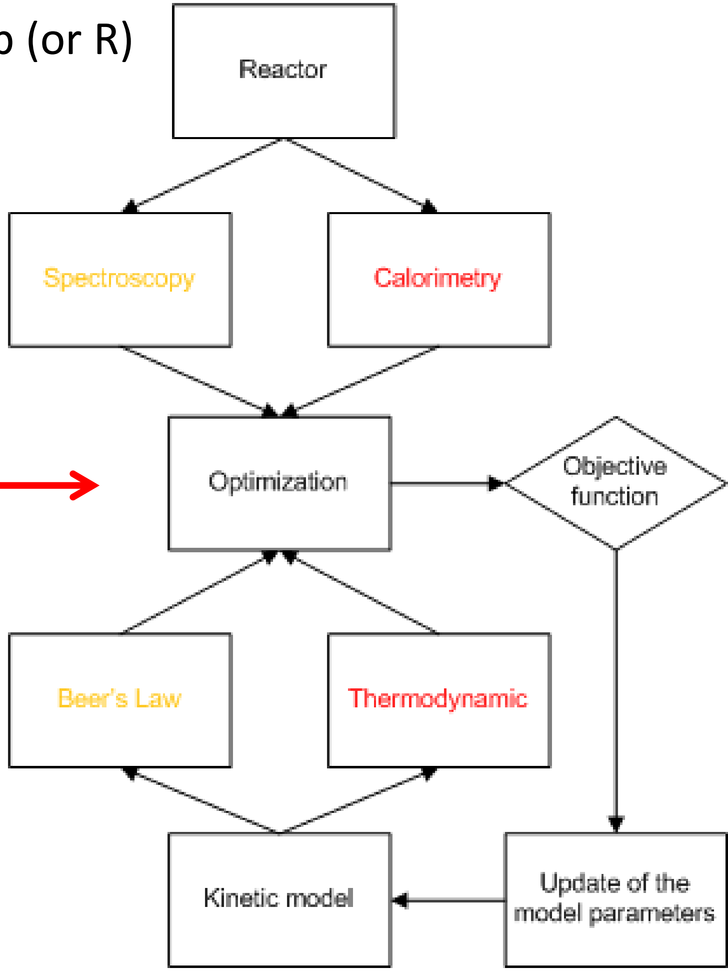
- to understand the underlying ideas and principles of the calibration-free modelling (fitting) of spectro-kinetic absorbance data
- to apply these principles to kinetic problems using Matlab (or R)

$Y_{\mathbf{Measured}}$

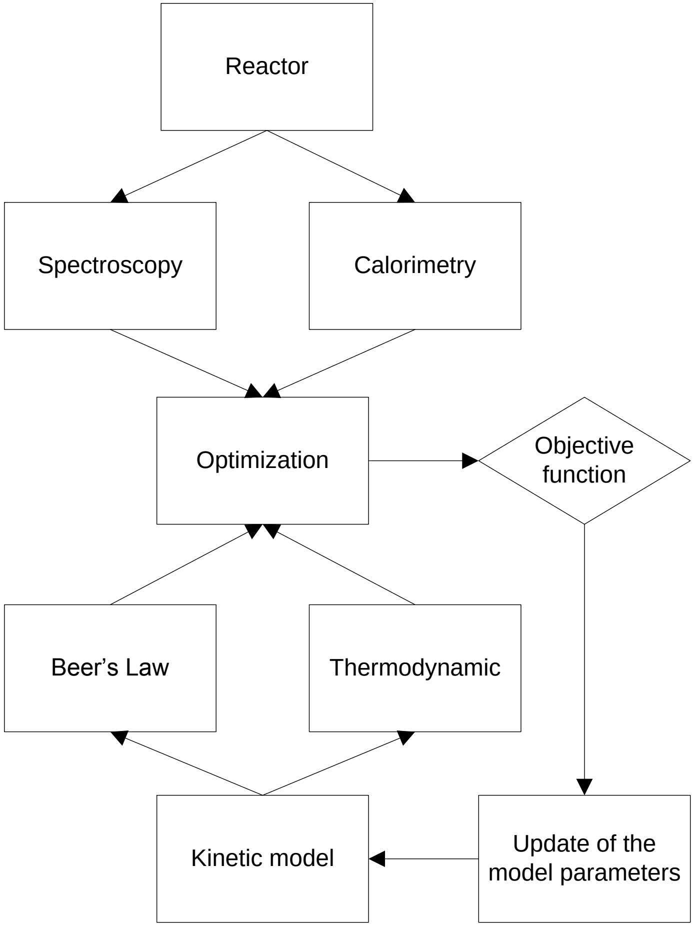
4c) $$\min_{k} \left\| Y_{\mathbf{Measured}} - C_{\mathbf{modelled}} A_{\mathbf{modelled}} \right\|^2$$

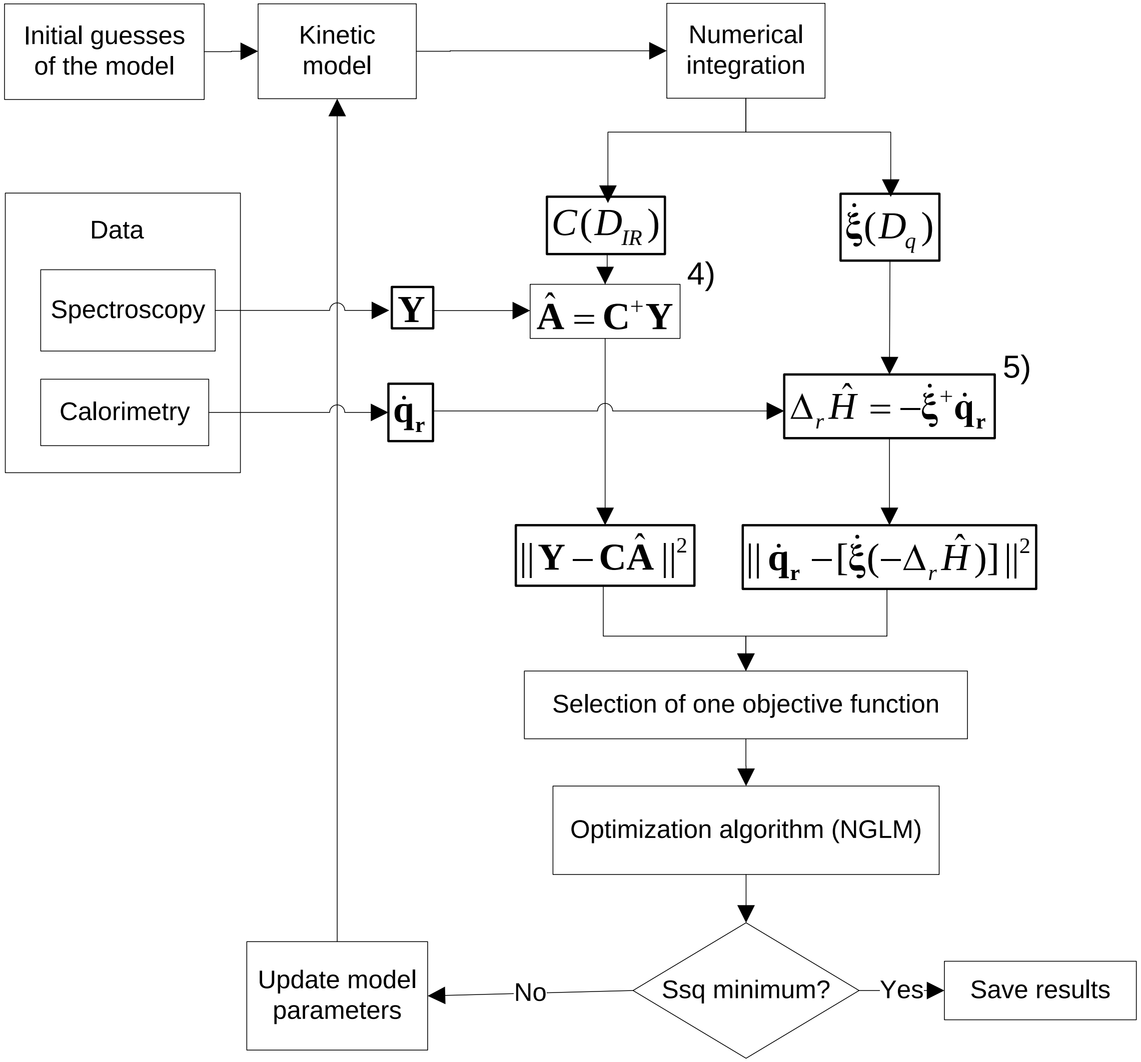
4b) $A_{\mathbf{modelled}}$ $$A_{\mathbf{modelled}} = C^{+}_{\mathbf{modelled}} Y_{\mathbf{Measured}}$$

4a) $C_{\mathbf{Modelled}}$ $$\begin{cases} \frac{dC_{Ac_2O}}{dt} & = r(k,t) + \frac{f_{dos}}{V_r(t)} \cdot (C_{dos,Ac_2O} - C_{Ac_2O}) \\ \frac{dC_{H_2O}}{dt} & = r(k,t) - \frac{f_{dos}}{V_r(t)} \cdot C_{H_2O} \\ \frac{dC_{AcOH}}{dt} & = -2 \cdot r(k,t) + \frac{f_{dos}}{V_r(t)} \cdot (C_{dos,AcOH} - C_{AcOH}) \\ \frac{dV_r}{dt} & = f_{dos} \end{cases}$$



4 Kinetic fitting of spectroscopic data with non-linear regression




Initial guesses of the model
Kinetic model
Numerical integration
$C(D_{IR})$
$\dot{\xi}(D_q)$
Data
Spectroscopy
Calorimetry
$\mathbf{Y}$
$\hat{\mathbf{A}} = \mathbf{C}^{+}\mathbf{Y}$
4)
$\dot{\mathbf{q}}_{\mathbf{r}}$
$\Delta_r \hat{H} = -\dot{\xi}^{+}\dot{\mathbf{q}}_{\mathbf{r}}$
5)
$\| \mathbf{Y} - \mathbf{C}\hat{\mathbf{A}} \|^2$
$\| \dot{\mathbf{q}}_{\mathbf{r}} - [\dot{\xi}(-\Delta_r \hat{H})] \|^2$
Selection of one objective function
Optimization algorithm (NGLM)
Ssq minimum?
No
Yes
Update model parameters
Save results

4c) Kinetic modelling – Fitting spectra, Y

```
% load spectro-kinetic experimental data (Y,t,w,c0,V0,F,cdos)
load('data_fit_CY.mat');
% Initial guess for the rate constant(s)
k0 = [7; 3]
% Fitting of Y (optimization of k0)
[k, ssq, Hessian, Ccalc, Acalc] =
               nglm_Y('r_Ycalc', k0, t, c0, Y, cdos, V0, F);
% Standard deviation on k
df    = prod(size(Y)) - length(k0) - prod(size(Acalc));
sig_r = sqrt(ssq/df)
sig_k = sig_r*sqrt(diag(inv(Hessian)))
```

$$A + B \xrightarrow{k_1} C$$

$$2C \xrightarrow{k_2} D$$

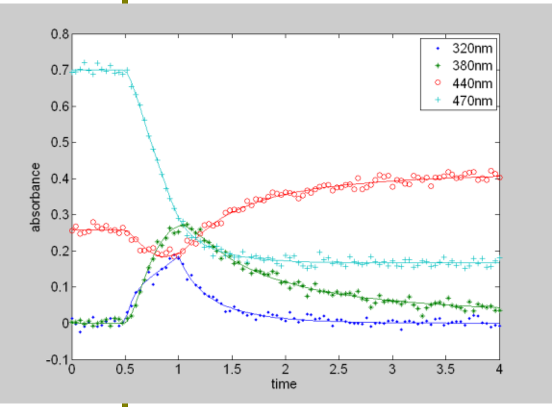
semi-batch
(*A* dosed)

Matlab output

```
k0 = 7 3

it=0, k(1)=7,       k(2)=3,       ssq=8.1585
it=1, k(1)=9.51204, k(2)=4.41167, ssq=7.97518
it=2, k(1)=10.404,  k(2)=4.84956, ssq=7.96041
it=3, k(1)=10.463,  k(2)=4.83658, ssq=7.9604

sig_r = 0.01
sig_k = 0.2664   0.1312
```

```
function [r, ssq, Ccalc, Acalc] = r_Ycalc(k, t, c0, Y, cdos, V0, F)
odeoptions   = odeset('RelTol', 1e-10, 'AbsTol', 1e-12);
[tdummy,CV] = ode45('ode_ApBtoC_2CtoD_semi', t, [c0 V0]',
                    odeoptions, k, F, t, cdos);
% Extraction of the concentration profiles
Ccalc        = CV(:,1:end-1);
Vcalc        = CV(:,end);
% Calculation of Acalc
Acalc        = Ccalc\Y;
% Calculation of the residuals
R            = Y - Ccalc*Acalc;
% Vectorization
r            = R(:);
% Calculation of the sum of squares
ssq          = sum(r.^2);
```

elimination of **A**



fitted vs 'measured' absorbances at selected wavelengths

4d) How to select a kinetic model

- Occam´s razor (*lex parsimoniae*)

  The principle states that among competing hypotheses, the one with the fewest assumptions should be selected. Other, more complicated solutions may ultimately prove correct, but—in the absence of certainty—the fewer assumptions that are made, the better.

- Akaike information criterion (AIC)

  The AIC is a measure of the relative quality of a statistical model for a given set of data. As such, AIC provides a means for model selection.

  $AIC = 2k - 2\ln(L)$

  where *k* is the number of parameters in the model, and *L* is the maximized value of the likelihood function for the model.

  Models with small AIC should be preferred.

- Bayesian information criterion (BIC)…and more

Concluding remarks on hard modelling section

- We “only” considered irreversible homogeneous solution chemistry & absorbance measurements but the principles can be extended to also deal with
  - instantaneous & kinetically observable equilibria
  - multiphasic transitions (e.g. surface catalysis, gas formation)
  - change of pH, ionic strength (activities), temperature
  - other data types (e.g. heat, pressure, pH, conductivity, particle size)

- Generally, this “only” requires an adaptation of the differential equations defining the mass transfer and a reassignment of the signal(s) to be fitted and/or parameters to be optimised

Framework

Chemometric will support you for:

- How to design/select the reactor/instrumentation?
- How to do the experimental design?
- How to “find” the correct kinetic model?
- What to “fit” and how ?
- How to determine the rate constant?
- How to “fit” if “C” can not be isolated and unknown?
- What about baseline drift, shift, noise level?
- Finally, are my fitted parameters “correct” to which extends?
- And we have to be quick to do all the above tasks

Reading material – Thank you

- **Factor Analysis in Chemistry**

  E.R. Malinowski, 3rd ed., Wiley, New York 2002

- **Practical Data Analysis in Chemistry**

  M. Maeder, Y.M. Neuhold, Elsevier, Amsterdam 2007

- **Practical Guide to Chemometrics**

  P. Gemperline (editor), 2nd ed., CRC Press, Boca Raton 2006

- **The Investigation of Organic Reactions and their Mechanisms**

  H. Maskill (editor), Blackwell Publishing, Oxford 2006

- **Evolving factor analysis for the resolution of overlapping chromatographic peaks**

  M. Maeder, Anal. Chem. 59 (1987), 527-530

- **Nonlinear Least-Squares Fitting of Multivariate Absorption Data**

  M. Maeder, A. Zuberbühler. Anal. Chem. 62 (1990), 2220-2224

- **Analyses of 3-way data from equilibrium and kinetic investigations**

  R. Dyson , M. Maeder, Y.M. Neuhold, G. Puxty. Anal. Chim. Act. 490 (2003), 99-108

- **Empirical kinetic modelling of on-line simultaneous infrared and calorimetric measurement using a pareto optimal approach and multi-objective genetic algorithm**

  S.I. Gianoli, G. Puxty, U. Fischer, M. Maeder, K. Hungerbühler. Chemom. Int. Lab. Syst. 85 (2007), 47-62

- **Tutorial on the Fitting of Kinetic Models to Multivariate Spectroscopic Measurements with Non-Linear Regression**

  G. Puxty, M. Maeder, K. Hungerbühler. Chemom. Int. Lab. Syst. 81 (2006), 149-164

- **Data Oriented Process Development: Determination of Reaction Parameters by Small-Scale Calorimetry with in situ Spectroscopy**

  G. Puxty, U. Fischer, M. Jecklin, K. Hungerbühler. Chimia 60 (2006), 605-610