Multivariate Resolution in Chemistry

Lecture 2

Roma Tauler
IIQAB-CSIC, Spain
e-mail: rtaqam@iiqab.csic.es

Lecture 2

- Resolution of two-way data.
- Resolution conditions.
 - Selective and pure variables.
 - Local rank
 - Natural constraints.
- Non-iterative and iterative resolution methods and algorithms.
- Multivariate Curve Resolution using Alternating Least Squares, MCR-ALS.
- Examples of application.

Multivariate (Soft) Self Modeling Curve Resolution (definition)

- Group of techniques which intend the **recovery of the response profiles** (spectra, pH profiles, time profiles, elution profiles,....) of more than one component in an **unresolved and unknown mixture** obtained from **chemical processes and systems** when **no (little) prior information is available** about the nature and/or composition of these mixtures.

Chemical reaction systems monitored using spectroscopic measurements

Analytical characterization of complex environmental, industrial and food mixtures using hyphenated methods (chromatography or continuous flow methods with spectroscopic detection).

Bilinearity!

Protein folding and dynamic protein-nucleic acid interaction processes.

Environmental source resolution and apportionment

source
distribution

source
composition

concn. of 96 organic compounds

$$d_{ij} = \sum_{k=1}^N c_{ik} s_{kj} + e_{ij}$$

Bilinearity!

Image analysis of spatially distributed chemicals on 2D surfaces measured using coupled microscopy-spectroscopy techniques in geological samples, biological tissues or food samples

Soft-modelling

MCR bilinear model for two way data:

$$d_{ij} = \sum_{n=1}^N c_{in} s_{nj} + e_{ij}$$

$$D = CS^T + E$$

d_{ij} is the data measurement (response) of variable j in sample i
 $n=1, \dots, N$ are the number of components (species, sources...)

c_{in} is the concentration of component n in sample i ;

s_{nj} is the response of component n
at variable j

Lecture 2

- Resolution of two-way data.
- Resolution conditions.
 - Selective and pure variables.
 - Local rank
 - Natural constraints.
- Non-iterative and iterative resolution methods and algorithms.
- Multivariate Curve Resolution using Alternating Least Squares, MCR-ALS.
- Examples of application.

Resolution conditions to reduce MCR rotation ambiguities (unique solutions?)

- Selective variables for every component
- Local rank conditions (Resolution Theorems)
- Natural Constraints
 - non-negativity
 - unimodality
 - closure (mass-balance)
- Multiway Data (i.e. trilinear data...)
- Hard-modelling constraints
 - mass-action law
 - rate law
 -
- Shape constraints (gaussian, lorentzian, assymmetric peak shape, log peak shape, ...)
-

Unique resolution conditions

First possibility: using **selective/pure** variables

elution time selective ranges,
where only one component is
present → spectra can be
estimated without
ambiguities

wavelength selective
Ranges, where only one
component absorbs →
elution profiles can be
estimated without
ambiguities

Detection of 'purest' (more selective) variables

Methods focused on finding the most representative (purest) rows (or columns) in a data matrix.

Based on PCA

- Key Set Factor Analysis (KSFA)

Based on the use of real variables

- Simple-to-use Interactive Self-modelling analysis (SIMPLISMA)
- Orthogonal Projection Approach (OPA)

How to detect purest/selective variables?

Selective variables are the more pure/representative/
dissimilar/orthogonal (linearly independent) variables..!

Examples of proposed methods for detection of selective variables:

- Key set variables KSFA E.D.Malinowski, Anal.Chim Acta, 134 (1982) 129; IKSFA, Chemolab, 6 (1989) 21
- SIMPLISMA: W.Windig & J.Guilmet, Anal. Chem., 63 (1991) 1425-1432)
- Orthogonal Projection Analysis OPA: F.Cuesta-Sanchez et al., Anal. Chem. 68 (1996) 79)
-

SIMPLISMA

- Finds the purest process or signal variables in a data set.

SIMPLISMA

HPLC-DAD Purest retention times

- **Variable purity**

$$p_i = \frac{s_i}{m_i}$$

s_i Std. deviation
 m_i Mean

Noisy variables

$$s_i \uparrow \quad m_i \downarrow \Rightarrow p_i \uparrow$$

SIMPLISMA

HPLC-DAD Purest retention times

- **Variable purity**

$$p_i = \frac{s_i}{m_i + f}$$

s_i Std. deviation

m_i Mean

f % noise (offset)

Noisy variables $\rightarrow p_i \downarrow$

SIMPLISMA

Working procedure

1. Selection of first pure variable. $\max(\mathbf{p}_i)$
2. Normalisation of spectra.
3. Selection of second pure variable.

- a. Calculation of weights (w_i)

$$\mathbf{w}_i = \mathbf{det}(\mathbf{Y}_i^T \mathbf{Y}_i)$$

- b. Recalculation of purity (p'_i)

$$\mathbf{p}'_i = \mathbf{w}_i \mathbf{p}_i$$

- c. Next purest variable. $\max(\mathbf{p}'_i)$

SIMPLISMA

Working procedure

3. Selection of third pure variable.

a. Calculation of weights (w_i)

$$w_i = \mathbf{det}(\mathbf{Y}_i^T \mathbf{Y}_i)$$

b. Recalculation of purity (p''_i)

$$p''_i = w_i p_i$$

c. Next purest variable. $\max(p''_i)$

.

.

.

SIMPLISMA

Graphical information

- **Purity spectrum.**

Plot of p_i vs. variables.

- **Std. deviation spectrum.**

Plot of 'purity corrected' std. dev. (cs_i) vs.
variables

$$CS_i = W_i S_i$$

SIMPLISMA

Graphical information

Concentration profiles

Mean spectrum

Std. deviation spectrum

1st pure spectrum

if 1st variable is too noisy
⇒ f is too low and should be
increased

SIMPLISMA

Graphical information

Concentration profiles

2nd pure spectrum

2nd std. dev. spectrum

SIMPLISMA

Graphical information

Concentration profiles

3rd pure spectrum

3rd std. dev. spectrum

SIMPLISMA

Graphical information

Concentration profiles

4th pure spectrum

4th std. dev. spectrum

SIMPLISMA

Graphical information

Noisy pattern in both spectra
No more significant contributions

SIMPLISMA

Information

- Purest variables in the two modes.
- Purest signal and concentration profiles.
- Number of compounds.

Unique resolution conditions

- Many chemical mixture systems (evolving or not) do not have selective variables for all the components of the system
- When selected variables are not (totally) selective, their detection is still very useful as an initial description of the system reducing its complexity and because they provide good initial estimations of species profiles useful for most of the resolution methods

Lecture 2

- Resolution of two-way data.
- Resolution conditions.
 - Selective and pure variables.
 - Local rank
 - Natural constraints.
- Non-iterative and iterative resolution methods and algorithms.
- Multivariate Curve Resolution using Alternating Least Squares, MCR-ALS.
- Examples of application.

Unique resolution conditions

Second possibility: using **local rank** information

What is local rank?

Local rank is the rank of reduced data regions in any of the two orders of the original data matrix

It can be obtained by **Evolving Factor Analysis** derived methods (EFA, FSMW-EFA, ...)

Conditions for unique solutions (unique resolution, uniqueness) based using local rank information have been described as: **Resolution Theorems**

Rolf Manne, On the resolution problem in hyphenated chromatography. Chemometrics and Intelligent Laboratory Systems, 1995, 27, 89-94

Resolution Theorems

Theorem 1: If all interfering compounds that appear inside the concentration window of a given analyte also appear outside this window, it is possible to calculate without ambiguities the concentration profile of the analyte

$$\mathbf{D}[\mathbf{I} - \mathbf{V}\mathbf{V}^T] = \mathbf{c}_a \left[\mathbf{s}_a^T - \sum_m (\mathbf{s}_a^T \mathbf{v}_m) \mathbf{v}_m^T \right]$$

$\mathbf{V}$ matrix defines the vector subspace where the analyte is not present and all the interferences are present. $\mathbf{V}$ matrix can be found by PCA (loadings) of the submatrix where the analyte is not present!

Resolution Theorems

This local rank information can be obtained from submatrix analysis (EFA, EFF)
 Matrix $\mathbf{V}^T$ may be obtained from PCA of the regions where the analyte
 is not present

**This is a rank
 one matrix!**

$$\mathbf{D}(\mathbf{I} - \mathbf{V}\mathbf{V}^T) = \mathbf{c}_a \left[\mathbf{s}_a^T - \sum_{m=1}^{n-1} (\mathbf{s}_a^T \mathbf{v}_m) \mathbf{v}_m^T \right]$$

concentration profile of analyte c_a may be resolved from $\mathbf{D}$ and $\mathbf{V}^T$

Resolution Theorems

Theorem 2: If for every interference the concentration window of the analyte has a subwindow where the interference is absent, then it is possible to calculate the spectrum of the analyte

**Local rank
information**

Resolution Theorems

Theorem 3. For a resolution based only upon rank information in the chromatographic direction the conditions of Theorems 1 and 2 are not only sufficient but also necessary conditions

Resolution based on local rank conditions

this system can
be totally resolved
using local rank
information!!!

this system cannot
be totally resolved
(only partially) based
only in local rank
information

Unique resolution conditions?

**In the case of
embedded peaks,
resolution
conditions based
on local rank are
not fulfilled! →**

**resolution without
ambiguities will be
difficult when a single
matrix is analyzed**

Conclusions about unique resolution conditions based on local rank analysis

In order to have a correct resolution of the system and to apply resolution theorems it is very important to have:

- 1) an accurate detection of local rank information → EFA based methods
- 2) This local rank information can be introduced in the resolution process using either:
 - ✓ non-iterative direct resolution methods
 - ✓ iterative optimization methods

Resolution Theorems

- Resolution theorems can be used in the two matrix directions (modes/orders), in the chromatographic and in the spectral direction.
- Resolution theorems can be easily extended to multiway data and augmented data matrices (unfolded, matricized three-way data) → Lecture 3
- Many resolution methods are implicitly based on these resolution theorems

Lecture 2

- Resolution of two-way data.
- Resolution conditions.
 - Selective and pure variables
 - Local rank
 - Natural constraints.
- Non-iterative and iterative resolution methods and algorithms.
- Multivariate Curve Resolution using Alternating Least Squares, MCR-ALS.
- Examples of application.

Unique resolution conditions

Third possibility: using **natural constraints**

- **Natural constraints** are previously known conditions that the profile solutions should have. We know that certain solutions are not correct!
- Even when non selective variables nor local rank resolutions conditions are present, **natural constraints** can be applied. They reduce significantly the number of possible solutions (rotation ambiguity)
- However, **natural constraints** alone, do not produce unique solutions in general

Natural constraints

- **Non negativity:**
 - species profiles in one or two orders are not negative (concentration and spectra profiles)
- **Unimodality:**
 - some species profiles have only one maximum (i.e. concentration profiles)
- **Closure**
 - the sum of species concentration is a known constant value (i.e. in reaction based systems = mass balance equation)

Non-negativity

**Constrained profile(s)
update
plain LS profile(s).**

Unimodality

Closure

C^*

Mass balance

C_c

$\rightarrow \Sigma = C_{\text{total}}$

Constraints

Concentration selectivity/local rank constraint

Hard-modelling

C^*

Physicochemical model

C_c

Hard + Soft modeling constraints

Unique resolution conditions

Forth possibility: by **multiway**, **multiset** data analysis and matrix **augmentation** strategies (**Lecture 3**)

- A set of correlated data matrices of the same system obtained under different conditions are simultaneously analyzed (**Matrix Augmentation**)
- Factor Analysis ambiguities can be solved more easily for three-way data, specially for **trilinear three-way** data

Lecture 2

- Resolution of two-way data.
- Resolution conditions.
 - Selective and pure variables
 - Local rank
 - Natural constraints.
- Non-iterative and iterative resolution methods and algorithms.
- Multivariate Curve Resolution using Alternating Least Squares, MCR-ALS.
- Examples of application.

Multivariate Curve Resolution (MCR) methods

•Non-iterative resolution methods

- ✓ Rank Annihilation Evolving Factor Analysis (RAEFA)
- ✓ Window Factor Analysis (WFA)
- ✓ Heuristic Evolving Latent Projections (HELP)
- ✓ Subwindow Factor Analysis (SFA)
- ✓ Gentle
- ✓

•Iterative resolution methods

- ✓ Iterative Factor Factor Analysis (ITF)
- ✓ Positive Matrix Factorization (PMF)
- ✓ **Alternating Least Squares (ALS)**
- ✓

Non-iterative resolution methods are mostly based on detection and use of local rank information

- Rank Annihilation by Evolving Factor Analysis (**RAEFA**, H.Gampp et al. Anal.Chim.Acta 193 (1987) 287)
- Non-iterative EFA (M.Maeder, Anal.Chem. 59 (1987) 527)
- **Window Factor Analysis** (**WFA**, E.R.Malinowski, J.Chemomet., 6 (1992) 29)
- Heuristic Evolving Latent Projections (**HELP**, O.M.Kvalheim et al., Anal.Chem. 64 (1992) 936)

WFA method description

E.R.Malinowski, J.Chemomet., 6 (1992) 29)

$$\mathbf{D} = \mathbf{C} \mathbf{S}^T = \sum c_i s_i^T \quad i=1, \dots, n$$

1. Evaluate the window where the analyte n is present (EFA, EFF..)
2. Create submatrix $\mathbf{D}^o$ deleting the window of the analyte n
3. Apply PCA to $\mathbf{D}^o = \mathbf{U}^o \mathbf{V}^{To} = \sum u_j^o v_j^{To} \quad j=1, \dots, m, m=n-1$
4. Spectra of the interferents are: $s_i = \sum \beta_{ij} v_j^{To} \quad j=1, \dots, m$
5. Spectra of the analyte lie in the orthogonal subspace of $\mathbf{V}^{To}$
6. Concentration of the analyte c_n can be calculated from:

$$(\mathbf{I} - \mathbf{V} \mathbf{V}^T) \mathbf{D} = \beta_{nn} s_n^o c_n = \mathbf{D}_n$$

c_n and s_n^o can be
obtained
directly!!

$\mathbf{D}_n$ is a rank one matrix
 s_n^o is part of the spectrum of the
analyte s_n which is orthogonal to
the interference spectra

Like 1st Resolution Theorem!!!

Non-iterative resolution methods based on detection and use of local rank information

Non-iterative resolution methods based on detection and use of local rank information

The main drawbacks of non-iterative resolution methods (like WFA) are:

- a) the impossibility to solve data sets with non-sequential profiles (e.g., data sets with embedded profiles)**
- b) the dangerous effects of a bad definition of concentration windows.**

Non-iterative resolution methods based on detection and use of local rank information

Improving WFA has been the main goal of modifications of this algorithm:

E.R. Malinowski, **Automatic Window Factor Analysis**. A more efficient method for determining concentration profiles from evolutionary spectra". J. Chemometr. 10, 273-279 (1996).

Subwindow Factor Analysis (SFA) based on the systematic comparison of matrix windows sharing one compound in common. R. Manne, H. Shen and Y. Liang. "Subwindow factor analysis". Chemom. Intell. Lab. Sys., 45, 171-176 (1999).

Iterative resolution methods (third alternative!)

Iterative Target Factor Analysis, ITTFA

- P.J. Gemperline, J.Chem.Inf.Comput.Sci., 1984, 24, 206-12
- B.G.M.Vandeginste et al., Anal.Chim.Acta 1985, 173, 253-264

Alternating Least Squares, ALS

- R.Tauler, A.Izquierdo-Ridorsa and E.Casassas. Chemometrics and Intelligent Laboratory Systems, 1993, 18, 293-300.
- R. Tauler, A.K. Smilde and B.R Kowalski. J. Chemometrics 1995, 9, 31-58.
- R.Tauler, Chemometrics and Intelligent Laboratory Systems, 1995, 30, 133-146.

Iterative Target Factor Analysis

a) Geometrical representation of ITTFA from initial needle targets x_{1in} and x_{2in}

b) Evolution of the shape of the two profiles through the ITTFA process

Iterative resolution methods

Iterative Target Factor Analysis ITTFA

ITTFA gets each concentration profile following the steps below:

1. Calculation of the score matrix by PCA.
2. Use of an estimated concentration profile as initial target.
3. Projection of the target onto the score space.
4. Constraint of the target projected.
5. Projection of the constrained target.
6. Go to 4 until convergence is achieved.

Lecture 2

- Resolution of two-way data.
- Resolution conditions.
 - Selective and pure variables
 - Local rank
 - Natural constraints.
- Non-iterative and iterative resolution methods and algorithms.
- Multivariate Curve Resolution using Alternating Least Squares, MCR-ALS.
- Examples of application.

Soft-modelling

MCR bilinear model for two way data:

$$d_{ij} = \sum_{n=1}^N c_{in} s_{nj} + e_{ij}$$

$$D = CS^T + E$$

d_{ij} is the data measurement (response) of variable j in sample i
 $n=1, \dots, N$ are the number of components (species, sources...)

c_{in} is the concentration of component n in sample i ;

s_{nj} is the response of component n
at variable j

Multivariate Curve Resolution (MCR)

Mixed information $\longrightarrow$ Pure component information

An algorithm to solve Bilinear models using
Multivariate Curve Resolution (MCR):
Alternating Least Squares (MCR-ALS)

$\mathbf{C}$ and $\mathbf{S}^T$ are obtained by solving
iteratively the two alternating LS equations:

$$\min_{\hat{\mathbf{C}}} \|\hat{\mathbf{D}}_{\text{PCA}} - \hat{\mathbf{C}}\hat{\mathbf{S}}^T\|$$

$$\min_{\mathbf{S}^T} \|\hat{\mathbf{D}}_{\text{PCA}} - \hat{\mathbf{C}}\hat{\mathbf{S}}^T\|$$

- Optional constraints (*local rank, non-negativity, unimodality, closure, ...*) are applied at each iteration
- Initial estimates of $\mathbf{C}$ or $\mathbf{S}$ are obtained from EFA or from pure variable detection methods.

Multivariate Curve Resolution

Alternating Least Squares

Model $\longrightarrow$

$$\mathbf{D} = \mathbf{CS}^T + \mathbf{E}$$
$$\hat{\mathbf{D}}_{\text{PCA}} = \mathbf{UV}^T$$

Algorithm to find
the Solution $\longrightarrow$

{

$$\min_{\hat{\mathbf{C}}, \text{constraints}} \left\| \hat{\mathbf{D}}_{\text{PCA}} - \mathbf{CS}^T \right\|$$

$$\min_{\mathbf{S}^T, \text{constraints}} \left\| \hat{\mathbf{D}}_{\text{PCA}} - \mathbf{CS}^T \right\|$$

Multivariate Curve Resolution Alternating Least Squares (MCR-ALS) Unconstrained Solution

- Initial estimates of $\mathbf{C}$ or $\mathbf{S}$ are obtained from EFA or from pure variable detection methods
- **Optional constraints are applied at each iteration !**

$$\mathbf{D} = \mathbf{C}\mathbf{S}^T + \mathbf{E}$$

$$1) \mathbf{S}^T = \mathbf{C}^+ \hat{\mathbf{D}}_{\text{PCA}}$$

$$2) \mathbf{C} = \hat{\mathbf{D}}_{\text{PCA}} (\mathbf{S}^T)^+$$

$\mathbf{C}^+$ and $(\mathbf{S}^T)^+$ are the pseudoinverses of $\mathbf{C}$ and $\mathbf{S}^T$ respectively

Matrix pseudoinverses

C and **S^T** are not square matrices. Their inverses are not defined

If they are full rank, i.e. the rank of **C** is equal to the number of its columns, and the rank of **S^T** is equal to the number of its rows,
The generalized inverse or pseudoinverse is defined:

$$\mathbf{D} = \mathbf{C} \mathbf{S}^T$$

$$\mathbf{C}^T \mathbf{D} = \mathbf{C}^T \mathbf{C} \mathbf{S}^T$$

$$(\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T \mathbf{D} = (\mathbf{C}^T \mathbf{C})^{-1} (\mathbf{C}^T \mathbf{C}) \mathbf{S}^T$$

$$(\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T \mathbf{D} = \mathbf{S}^T$$

$$\mathbf{C}^+ \mathbf{D} = \mathbf{S}^T$$

$$\text{Where } \mathbf{C}^+ = (\mathbf{C}^T \mathbf{C})^{-1} \mathbf{C}^T$$

$$\mathbf{D} = \mathbf{C} \mathbf{S}^T$$

$$\mathbf{D} \mathbf{S} = \mathbf{C} \mathbf{S}^T \mathbf{S}$$

$$\mathbf{D} \mathbf{S} (\mathbf{S}^T \mathbf{S})^{-1} = \mathbf{C} (\mathbf{S}^T \mathbf{S}) (\mathbf{S}^T \mathbf{S})^{-1}$$

$$\mathbf{D} \mathbf{S} (\mathbf{S}^T \mathbf{S})^{-1} = \mathbf{C}$$

$$\mathbf{D} (\mathbf{S}^T)^+ = \mathbf{C}$$

$$\text{Where } (\mathbf{S}^T)^+ = \mathbf{S} (\mathbf{S}^T \mathbf{S})^{-1}$$

C⁺ and **(S^T)⁺** are the pseudoinverses of **C** and **S^T** respectively. They also provide the best least squares estimations of the overdetermined linear system of equations. If **C** and **S^T** are not full rank, it is still possible to define their pseudoinverses using SVD

Flowchart of MCR-ALS

Iterative resolution methods

Alternating Least Squares MCR-ALS

ALS optimizes concentration and spectra profiles using a constrained alternating least squares method. The main steps of the method are:

1. Calculation of the PCA reproduced data matrix.
2. Calculation of initial estimations of concentration or spectral profiles (e.g, using SIMPLISMA or EFA).
3. Alternating Least Squares
 - Iterative least squares constrained estimation of **C** or **S^T**
 - Iterative least squares constrained estimation of **S^T** or **C**
 - Test convergence
4. Interpretation of results

Flowchart of MCR-ALS

Journal of Chemometrics, 1995, 9, 31-58; Chemomet.Intel. Lab. Systems, 1995, 30, 133-146

Journal of Chemometrics, 2001, 15, 749-7; Analytica Chimica Acta, 2003, 500, 195-210

Until recently

MCR-ALS input had to be typed in
the MATLAB command line

Troublesome and difficult in complex cases where several data
matrices are simultaneously analyzed and/or different constraints
are applied to each of them for an optimal resolution

Now

**A graphical user-friendly
interface for MCR-ALS**

J. Jaumot, R. Gargallo, A. de Juan and R. Tauler, **Chemometrics
and Intelligent Laboratory Systems**, 2005, 76(1) 101-110

*Multivariate Curve Resolution
Home Page*

<http://www.ub.es/gesq/mcr/mcr.htm>

Example. Analysis of multiple experiments. Analysis of 4 HPLC-DAD runs each of them containing four compounds

Selection of ALS constraints

No-negativity

☒ Yes?

☐ Conc

☐ Spectra

☒ Conc & Spec

Implementation for conc

fnrls

Implementation for spec

fnrls

Nr. of species with non-neg conc

2

Nr. of species with non-neg spec

2

Enter a vector of positive profiles

Enter a vector of positive profiles

Unimodality

☒ Yes?

☒ Conc

☐ Spectra

☐ Conc & Spec

Implementation of the unimodality constraint

average

Nr. of species with unimodal conc

2

Nr. of species with unimodal spec

select...

Species with unimodal conc?

Species with unimodal spec?

Constraint tolerance for conc

1.3

Constraint tolerance for spec

Closure

☒ Yes?

☒ Conc

☐ Spectra

☐ Closure variable?

Nr. of closure constraints to be included?

1

First Closure constant Equal to

1

Second Closure constant Equal to

First variable closure constants

Second variable closure constants

Closure condition

lower or equal than

Closure condition

select...

Which species are in 1st closure?

☒ All?

1 1

Which species are in 2nd closure?

☐ All?

Equality constraints in conc profiles

☐ Yes?

Select csel matrix:

Constraints are

select...

Equality constraints in spectra profiles

☐ Yes?

Select ssl matrix:

Constraints are

select...

Optimization parameters

Nr. of iterations

500

Convergence criterion

0.1

☒ Graphical output

Output

Concentration

copt

Std. dev.

Area opt

Spectra

sopt

Residuals

Ratio opt

Optimize

Done

Cancel

Alternating Least Squares

Initial estimates

- from EFA derived methods (for evolving methods like chromatography, titrations...)
- from 'pure' variable (SIMPLISMA) detection methods (for non-evolving methods and/or for very poorly resolved systems...)
- from individually and directly selected from the data using chemical reasoning (i.e first and last spectrum; isosbestic points,)
- from known profiles ...

Alternating Least Squares with constraints

- **Natural constraints:** non-negativity; unimodality, closure,...
- **Equality constraints:** selectivity, zero concentration windows, known profiles...
- **Optional Shape constraints** (gaussian shapes, asymmetric shapes)
- **Hard modeling constraints** (rate law, equilibrium mass-action law...)
-

How to implement constrained ALS optimization algorithms in optimal way from a least squares sense?

Considerations:

How to implement these algorithms in a way that all the constraints be fulfilled simultaneously at the same time (in every least squares step - in one LS shot- of the optimization)?

Updating (substitution) methods do work well most of the times!
Why? Because the optimal solutions which better fit the data (apart from noise and degrees of freedom) **do also fulfill** the constraints of the system

Constraints are used **to lead** the optimization in the right direction within feasible band solutions. .

Implementation of constraints

Non-negativity constraints case

- a) forcing values during iteration (e.g negative values to zero)
 - ✓ intuitive
 - ✓ fast
 - ✓ easy to implement
 - ✓ it can be used individually for each profile independently
 - ✓ less efficient
- b) using non-negative rigorous least squares optimization procedures:
 - ✓ more statistically efficient
 - ✓ more efficient
 - ✓ more difficult to implement
 - ✓ it has to be used to all profiles simultaneously
 - ✓ different approaches (penalty functions, constrained optimization, elimination...

How to implement constrained ALS optimization algorithms in optimal way from a least squares sense?

Different rigorous least-squares approaches have been proposed

- **Non-negative least squares methods** (Lawson CL, Hanson RJ. Solving Least Squares Problems. Prentice-Hall: 1974; Bro R, de Jong S. J. Chemometrics 1997; 11: 393–40; Mark H. Van Benthem and Michael R. Keenan, Journal of Chemometrics, 18, 441-450; ...)
- **Unimodal least-squares** approaches (R. Bro, N.D. Sidiropoulos, J. of Chemometrics, 1998, 12, 223-247)
- **Equality constraints** (Van Benthem M, Keenan M, Haaland D. J. Chemometrics 2002; 16, 613–622....)
- Use of **penalty terms in the objective functions to optimize**
- **Non-linear optimization with non-linear constraints** (PMF, Multilinear Engine, sequential quadratic programming.....)

Are still active the constraints at the optimum ALS solution?

Checking active constraints:

ALS solutions $\mathbf{D}_{\text{PCA}}$, $\mathbf{C}_{\text{ALS}}$, $\mathbf{S}_{\text{ALS}}$

New **unconstrained** solutions

$$\mathbf{C}_{\text{unc}} = \mathbf{D}_{\text{PCA}} (\mathbf{S}_{\text{ALS}}^T)^+$$

$$\mathbf{S}_{\text{unc}}^T = (\mathbf{C}_{\text{ALS}})^+ \mathbf{D}_{\text{PCA}}$$

Active non-negativity constraints:

$\mathbf{C}$ matrix

r	c	value
19	1	-4.1408e-003
21	1	-3.2580e-003
23	1	-1.8209e-003
24	1	-3.3004e-003
1	2	-1.1663e-002
2	2	-2.1166e-002
3	2	-2.1081e-002
4	2	-3.8524e-003
25	2	-1.9865e-003
26	2	-1.3210e-003
7	3	-5.9754e-003
8	3	-5.5289e-004

Deviations
are small!!!

$\mathbf{S}^T$ matrix

Empty matrix: 0-by-3

**Proposal: Check ALS
solutions for active
constraints and if
deviations are large!**

Implementation of unimodality constraints

➤ ‘vertical’ unimodality: forcing non-unimodal parts of the profile to zero

➤ ‘horizontal’ unimodality: forcing non-unimodal parts of the profile to be equal to the last unimodal value

➤ ‘average’ unimodality: forcing non-unimodal parts of the profile to be an average between the two extreme values being still unimodal

➤ using **monotone regression** procedures

Implementation of closure/ /normalization constraints

Equality constraints:

Closure constraints

experimental point i, 3 concn profiles

$$c_{i1} + c_{i2} + c_{i3} = t_i$$

$$c_{i1}r_1 + c_{i2}r_2 + c_{i3}r_3 = t_i$$

$$\mathbf{C} \mathbf{r} = \mathbf{t}$$

$$\mathbf{r} = \mathbf{C}^+ \mathbf{t}$$

Normalization constraints

$\max(\mathbf{s}) = 1$, spectra maximum

$\max(\mathbf{c}) = 1$, peak maximum

$\|(\mathbf{s})\| = 1$, area, length,...

.....

These are equality constraints!

Implementation of selectivity/local rank constraints

Using a masking $\mathbf{C}_{\text{sel}}$ or $\mathbf{S}_{\text{sel}}^T$ matrix

$$\mathbf{C}_{\text{sel}} = \begin{pmatrix} x & 0 & 0 \\ x & x & 0 \\ x & x & x \\ \dots & \dots & \dots \\ x & x & x \\ 0 & x & x \\ 0 & x & x \end{pmatrix}$$

From local rank (EFA) setting some values to zero

$$\mathbf{S}_{\text{sel}}^T = \begin{bmatrix} x & x & x & \dots & x & x & x & x & x \\ k & k & k & \dots & k & k & k & & \\ x & x & x & \dots & x & x & x & & \end{bmatrix}$$

Fixing a known spectrum

Solving intensity ambiguities in MCR-ALS

$$d_{ij} = \sum_n c_{in} s_{nj} = \sum_n k c_{in} \frac{1}{k} s_{nj}$$

k is arbitrary. How to find the right one?

**In the simultaneous analysis of multiple data matrices
intensity/scale ambiguities can be solved**

a) in relative terms (directly)

b) in absolute terms using external knowledge

Two-way data

MCR-ALS for quantitative determinations

Talanta, 2008, 74, 1201-10

Validation of the quantitative determination: spectrophotometric analysis of nucleic bases mixtures

MCR-ALS analyte concentration prediction

analyte	RMSEP	SEP	bias	RE %	slope	offset	corcoef	sdres
1.0000	0.0041	0.0042	0.0003	1.7714	1.0044	-0.0023	0.9997	0.0179
2.0000	0.0050	0.0050	-0.0012	1.4081	1.0019	0.0002	0.9998	0.0217
3.0000	0.0117	0.0120	-0.0001	2.8427	1.0092	-0.0052	0.9992	0.0511
4.0000	0.0046	0.0047	0.0007	1.7384	1.0010	-0.0011	0.9999	0.0205

MCR-ALS spectra prediction

In parenthesis correlation (similarity) respect initial estimates
red predicted, blue 'true'

Protein and moisture determination in agricultural samples (ray-grass) by PLSR and MCR-ALS

Talanta, 2008, 74, 1201-10

Example of application: NIR real data

Calibration data set
86 samples
87 wavelengths
no-pretreatment

Validation data set
31 samples
87 wavelengths
no-pretreatment

Determination of moisture and protein

Moisture concentrations
in the calibration and
validation data sets

Protein concentrations
in the calibration and
validation data sets

	RMSEP		SEP		Bias		Correlation		RE (%)	
	<i>ALS</i>	<i>PLS</i>	<i>ALS</i>	<i>PLS</i>	<i>ALS</i>	<i>PLS</i>	<i>ALS</i>	<i>PLS</i>	<i>ALS</i>	<i>PLS</i>
HUM	0.312	0.249	0.315	0.248	7.30 e^{-4}	4.50 e^{-2}	0.9755	0.986	3.70	2.96
PB	0.782	0.564	0.788	0.571	7.35 e^{-2}	3.31 e^{-2}	0.9860	0.993	4.65	3.67

Soft-Hard modelling

- All or some of the concentration profiles can be constrained.
- All or some of the batches can be constrained.

Implementation of hard modelling and shape constraints

$$D = C S^T$$

$$\begin{aligned} \min ||D - C S^T|| \\ \text{ALS } (D, S^T) \rightarrow C \\ \text{ALS } (D, C) \rightarrow S^T \end{aligned}$$

C_{soft}

**rate
Law**

$C_{\text{soft/hard}}$

Ordinary differential equations

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

$$[B] = [A]_0 \frac{k_1}{k_1 - k_2} (e^{-k_1 t} - e^{-k_2 t})$$

.....

.....

Integration

$$\frac{d[A]}{dt} = -k_1 [A]$$

$$\frac{d[B]}{dt} = k_1 [A] - k_2 [B]$$

.....

.....

Quality of MCR Solutions

Rotational Ambiguities

Factor Analysis (PCA) Data Matrix Decomposition

$$\mathbf{D} = \mathbf{U} \mathbf{V}^T + \mathbf{E}$$

‘True’ Data Matrix Decomposition

$$\mathbf{D} = \mathbf{C} \mathbf{S}^T + \mathbf{E}$$

$$\mathbf{D} = \mathbf{U} \mathbf{T} \mathbf{T}^{-1} \mathbf{V}^T + \mathbf{E} = \mathbf{C} \mathbf{S}^T + \mathbf{E}$$

$$\mathbf{C} = \mathbf{U} \mathbf{T}; \quad \mathbf{S}^T = \mathbf{T}^{-1} \mathbf{V}^T$$

How to find the rotation matrix $\mathbf{T}$?

Matrix decomposition is not unique!

$\mathbf{T}(\mathbf{N},\mathbf{N})$ is any non-singular matrix

There is rotational freedom for $\mathbf{T}$

It is possible to define bands and limits for the feasible solutions (T_{max} y T_{min})?

How T_{max} and T_{min} can be calculated from the constraints of the system

Constrained Non-Linear Optimization Problem (NCP)

Find $\mathbf{T}$ which makes: $\min/\max f(\mathbf{T})$
 under $g_e(\mathbf{T}) = 0$
 and $g_i(\mathbf{T}) \leq 0$

where $\mathbf{T}$ is the matrix of variables, $f(\mathbf{T})$ is a scalar non-linear function of $\mathbf{T}$ and $\mathbf{g}(\mathbf{T})$ is the vector of non-linear constraints

Matlab Optimizarion Toolbox *fmincon* function

1) What are the variables of the problem?

$\mathbf{T}$ (rotation matrix),
 $\mathbf{D} = \mathbf{C} \mathbf{T} \mathbf{T}^T \mathbf{S}^T$

2) What is the objective function $f(\mathbf{T})$ to optimize?

For every species $i = 1, \dots, ns$

$$f_i(\mathbf{T}) = \frac{\|c_i s_i\|}{\|\mathbf{C} \mathbf{S}^T\|} \quad \text{or} \quad f_i(\mathbf{T}) = \frac{\sum_j c_{ij} s_{ij}}{\sum_{i,j} c_{ij} s_{ij}}$$

$f(\mathbf{T})$ is a scalar value between 0 and 1!

This function gives the relative contribution of species i compared to the global measured signal!

3) What are the constraints $g(\mathbf{T})$?

The following constraints are considered

normalization/closure	$g_{\text{norm}}/g_{\text{clos}}$
non-negativity	$g_{\text{cneg}}/g_{\text{sneg}}$
known values/selectivity	$g_{\text{known}}/g_{\text{sel}}$
unimodality	g_{unim}
trilinearity (three-way data)	g_{tril}

Are they equality or inequality constraints?

4) What are the initial estimations of $\mathbf{C}$ and $\mathbf{S}^T$?

- Initial estimations of $\mathbf{C}$ y $\mathbf{S}^T$ are obtained by MCR-ALS
- Initial estimations should fulfill the constraints of the system (*non-negativity, unimodality, closure, selectivity, local rank, ...*)

5) What are the initial values of $\mathbf{T}$?

- NCP depends on initial values of $\mathbf{T}$! (local minima, convergence, speed ...)

$$\mathbf{T}_{\text{ini}} = \text{eye}(\mathbf{N}) = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 1 \end{pmatrix}$$

Optimization algorithm

R.Tauler. **Journal of Chemometrics**, 2001, 15, 627-646

Calculation of feasible bands in the resolution of a single chromatographic run (run 1)

Applied constraints were spectra and elution profiles non-negativity and spectra normalization:

elution profiles

spectra profiles

Calculation of feasible bands in the resolution of a single chromatographic run (run 1)

Applied constraints were spectra and elution profiles non-negativity, spectra normalization:, and **unimodality**

Calculation of feasible bands in the resolution of a single chromatographic run (run 1)

Applied constraints were spectra and elution profiles non-negativity, spectra normalization:, and **selectivity/local rank** (31-51, 45-51, 1-8,1-15)

Evaluation of boundaries of feasible bands:

Previous studies

- W.H.Lawton and E.A.Sylvestre, *Technometrics*, 1971, 13, 617-633
- O.S.Borgen and B.R.Kowalski, *Anal. Chim. Acta*, 1985, 174, 1-26
- K.Kasaki, S.Kawata, S.Minami, *Appl. Opt.*, 1983 (22), 3599-3603
- R.C.Henry and B.M.Kim (*Chemomet. and Intell. Lab. Syst.*, 1990, 8, 205-216)
- P.D.Wentzell, J-H. Wang, L.F.Loucks and K.M.Miller (*Can.J.Chem.* 76, 1144-1155 (1998))
- P. Gemperline (*Analytical Chemistry*, 1999, 71, 5398-5404)
- R.Tauler (*J.of Chemometrics* 2001, 15, 627-46)
- M.Legger and P.D.Wentzell, *Chemomet and Intell. Lab. Syst.*, 2002, 171-188

Quality of MCR results

Error propagation and resampling methods

- How **experimental error/noise** in the input data matrices affects MCR-ALS results?
 - For **ALS** calculations there is **no known analytical formula** to calculate error estimations. (i.e. like in linear least-squares regressions)
 - Bootstrap estimations using resampling methods is attempted

MCR-ALS: Quality Assessment

Propagation of experimental noise into the MCR-ALS solutions

Experimental noise is propagated into the MCR-ALS solutions and causes **uncertainties** in the obtained results.

To estimate these uncertainties for non-linear models like MCR-ALS computer intensive **resampling** methods can be used

Mean, max and min profiles Confidence range profiles

(J. of Chemometrics, 2004, 18, 327–340; J.Chemometrics, 2006, 20, 4-67)

Error Propagation

Parameter Confidence Range

		Real		0.1 %		1 %		2 %		5 %	
		pk1	pk2	pk1	pk2	pk1	pk2	pk1	pk2	pk1	pk2
Theoretical Value	Value	3.666 0	4.924 4	-	-	-	-	-	-	-	-
MonteCarlo Simulations	Value	-	-	3.666	4.924	3.669	4.926	3.676	4.917	3.976	5.074
	Stand. dev.	-	-	0.001	0.001	0.006 5	0.012	0.012	0.024	0.434	0.759
Noise Addition	Value	-	-	3.654	4.922	3.659	4.913	3.665	4.910	4.075	5.330
	Stand. dev.	-	-	0.001	0.002	0.006	0.026	0.010	0.040	0.487	1.122
JackKnife	Value	-	-	3.655	4.920	3.660	4.913	3.667	4.913	4.082	5.329
	Stand. dev.	-	-	0.004	0.003	0.009	0.024	0.012	0.047	0.514	1.091

Maximum Likelihood MCR-ALS solutions

$$Q = D - \hat{C}_{ALS} \hat{S}_{ALS}^T, \quad \frac{\partial Q^2}{\partial S^T} = 0, \quad \frac{\partial Q^2}{\partial C} = 0$$

Without including
uncertainties

$$Q^2 = \sum_{i=1}^m \sum_{j=1}^n (d_{i,j} - \hat{d}_{i,j})^2$$

Unconstrained ALS solution

$$S^T = (C^T C)^{-1} C \hat{D}_{PCA} = C^+ \hat{D}_{PCA}$$

$$C = \hat{D}_{PCA} S (S^T S)^{-1} = \hat{D}_{PCA} (S^T)^+$$

Including uncertainties

$\sigma_{i,j}$

$$Q^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(d_{i,j} - \hat{d}_{i,j})^2}{\sigma_{i,j}^2}$$

Unconstrained WALS solution

$$\mathbf{W}_i = \Sigma_i^{-1}, \quad \text{rows or} \quad \Sigma = \{\sigma_{i,j}\}$$

$$\mathbf{W}_j = \Sigma_j^{-1}, \quad \text{columns}$$

$$c(i,:) = d(i,:) \mathbf{W}_i S (S^T \mathbf{W}_i S)^{-1}$$

$$s^T(:,j) = (C^T \mathbf{W}_j C)^{-1} C^T \mathbf{W}_j d(:,j);$$

MCR-ALS results quality assesment

Data Fitting

- lof %

$$lof = 100 \sqrt{\frac{\sum_{i=1}^n \sum_{j=1}^m e_{i,j}^2}{\sum_{i=1}^n \sum_{j=1}^m x_{i,j}^2}}, \quad e_{i,j} = x_{i,j} - \hat{x}_{i,j}$$

- R %

$$R^2 = 100 \frac{\sum_{i=1}^n \sum_{j=1}^m x_{i,j}^2 - \sum_{i=1}^n \sum_{j=1}^m e_{i,j}^2}{\sum_{i=1}^n \sum_{j=1}^m x_{i,j}^2}$$

Profiles recovery

- r^2 (similarity)

$$r^2 = \cos \alpha = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

- recovery angles measured by the inverse cosine α , expressed in hexadecimal degrees

$$\alpha = a \cos d(r^2)$$

r^2	1	0.99	0.95	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.00
α	0	8.1	18	26	37	46	53	60	66	72	78	84	90

lof (%) = 14%
 R^2 98.0%
 mean(S/N)=21.7

**HOMOCEDASTIC
 NOISE CASE**

Noise structure:

$$r = 0.01 * \max(\max(Y)) = 3.21$$

$$S = I . * r$$

$$E = S . * N(0,1)$$

G

FT

815.2
 346.6
 104.1
 62.9
 0.0

SVD
Y E X

818.1
 348.9
 112.9
 66.1
 37.0

39.4
 36.6

No noise and homocedastic noise cases results

recovery angles α

System	init	method	lof %	R2%	f1 g1	f2 g2	f3 g3	f4 g4
No noise	true	ALS	0	100	0	0	0	0
					0	0	0	0
No noise	purest	ALS	0	100	1.8	11	7.9	5.0
					5.9	9.1	13	2.8
max band	-	Bands	0	100	3.1	13	7.5	5.5
					8.2	18	10	1.7
min band	-	Bands	0	100	2.1	3.7	3.9	3.9
					5.2	8.1	14	3.0
Homo noise	true	ALS	12.6	98.4	3.0	12	8.7	2.1
					4.8	12	9.0	2.4
Homo noise	purest	ALS	12.6	98.4	3.0	17	8.5	5.0
					7.1	12	16	3.7
Homo noise	-----	Theor	14.0	98.0	----	----	----	----
Homo noise	-----	PCA	12.6	98.4	----	----	----	----

lof (%) = 12, 25, 44%
 R^2 99, 94, 80%
 mean(S/N) = 17, 10, 3

HETEROCEDASTIC NOISE CASE Low, Medium, High

G

F^T

Noise structure:

$r = 5, 10, 20$

$S = r \cdot R(0,1)$ (interv 0-1)

$E = S \cdot N(0,1)$ Normal Distributed

random numbers

SVD

weighting
recovery

overall
improvement

Hoterocedastic noise case results

System (Case)	init	w	lof % exp	R ² % exp	f1 g1	recovery angles α			
						f2 g2	f3 g3	f4 g4	
Hetero noise (low)	purest	ALS	10.7	98.8	3.1	14	9.0	3.8	
					7.0	10	15	4.3	
Hetero noise (low)	purest	WALS	12.0	98.6	2.6	12	15	4.3	
					7.8	15	15	3.7	
Theoretical PCA	----	----	12.0	98.6	----	----	----	----	
	----	----	10.7	98.8	----	----	----	----	
Hetero noise (medium)	purest	ALS	22.3	95.0	7.7	22	22	5.7	
					7.2	21	24	4.5	
Hetero noise (medium)	purest	WALS	24.0	94.2	6.6	22	18	5.7	
					7.4	14	17	5.5	
Theoretical PCA	----	----	25.0	93.6	----	----	----	----	
	----	----	22.0	95.1	----	----	----	----	
Hetero noise (high)	purest	ALS	40.0	84.0	12	33	38	10	
					15	38	34	9.0	
Hetero noise (high)	purest	WALS	43.1	81.4	12	26	25	6.0	
					5.0	27	16	3.0	
Theoretical PCA	----	----	44.2	80.4	----	----	----	----	
	----	----	40.8	83.4	----	----	----	----	

Lecture 2

- Resolution of two-way data.
- Resolution conditions.
 - Selective and pure variables
 - Local rank
 - Natural constraints.
- Non-iterative and iterative resolution methods and algorithms.
- Multivariate Curve Resolution using Alternating Least Squares, MCR-ALS.
- Examples of application.

Spectrometric titrations: An easy way for the generation of two- and three-way data in the study of chemical reactions and interactions

Three spectrometric titrations of a complexation system at different ligand to metal ratios R

MCR-ALS resolved concentration profiles at $R=1.5$

MCR-ALS resolved concentration profiles at R=2.0

MCR-ALS resolved concentration profiles at R=3.0

MCR-ALS resolved spectra profiles

Process analysis

One process IR run (raw data)

2nd derivative

2nd derivative
and PCA
(3 PCs)

R. Tauler, B. Kowalski and S. Fleming Anal. Chem., 65 (1993) 2040-47

EFA of 2nd derivative data:
initial estimation of process profiles
for 3 components

ALS resolved pure IR spectra profiles

ALS resolved pure concentration profiles
in the simultaneous analysis of eighth
runs of the process

Study of conformational equilibria of polynucleotides

poly(adenylic)-poly(uridylic) acid system Melting data

*R. Tauler, R. Gargallo, M. Vives and
A. Izquierdo-Ridorsa*

*Chemometrics and Intelligent Lab
Systems, 1998*

poly(A)-poly(U) ds **poly(A)-poly(U)-poly(U) ts**

Concentration of 96 organic compounds in 22 samples of Northwestern Mediterranean Sea

source contribution profiles using

resolved composition profiles using nnls

Historical Evolution of Multivariate Curve Resolution Methods

- Extension to more than two components
- Target Factor Analysis and Iterative Target Factor Analysis Methods
- Local Rank Detection, Evolving Factor Analysis, Window Factor Analysis.
- Rank Annihilation derived methods
- Detection and selection of pure (selective) variables based methods
- Alternating Least Squares methods, 1992
- Implementation of soft modelling constraints (non-negativity, unimodality, closure, selectivity, local rank,...) 1993
- Extension to higher order data, multiway methods (extension of bilinear models to augmented data matrices), 1993-5
- Trilinear (PARAFAC) models, 1997
- Implementation of hard-modelling constraints, 1997
- Breaking rank deficiencies by matrix augmentation, 1998
- Calculation of feasible bands, 2001
- Noise propagation, 2002
- Tucker models, 2005
- Weighted Alternating Least Squares method (Maximum Likelihood), 2006
- ...