Multivariate Curve Resolution: theory and applications

Romà Tauler, Abril, 2005
Multivariate Curve Resolution

Mixed information → Pure component information

D → C → ST

Pure concentration profiles
Chemical model
Process evolution
Compound contribution

Pure signals
Compound identity

Retention times
Wavelengths

Absorbance
Flowchart of MCR-ALS

\[ D = C S^T + E \]
(bilinear model)

- Data Matrix decomposition according to a bilinear model
- Estimation of the number of components
- Initial Estimation
- ALS optimization
- Results of the ALS optimization procedure: Fit and Diagnostics

\[ \min_C \| \hat{D}_{PCA} - \hat{C} \hat{S}^T \| \]
\[ \min_{\hat{S}^T} \| D_{PCA} - \hat{C} \hat{S}^T \| \]

Constraints

Non-negativity

Unimodality

Concentration profiles
Chromatographic peaks
Voltammograms

Concentration selectivity/local rank constraint

Mass balance or Closure constraint

We know that this region is not rank 3, but rank 2!
Hard + soft modelling constraints
MCR-ALS hybrid (grey) models

Hard modeling constraints

Hard modelling examples: mass action law and rate laws

- Kinetic processes
- Equilibrium processes

Physicochemical model

Hard + Soft modeling constraints

- Soft model (non-negativity)
- Physicochemical model (mass action law, rate law)
- Kinetic processes
- Equilibrium processes
Multiway data: Multiple measurement orders/modes/directions/ways

Example: Three-way data

Regular data cube

Other three-way arrangements

A series of two-way data sets with common information in one or more modes.
Data augmentations in MCR

The same experiment monitored with different techniques

\[ D \times C = S_1^T S_2^T S_3^T \]

Row-wise

Column-wise

Row and column-wise

Several experiments monitored with the same technique

Several experiments monitored with several techniques
Trilinearity Constraint (flexible to every species)  
Extension of MCR-ALS to multilinear systems

\[
\begin{bmatrix}
D_1 \\
D_2 \\
D_3
\end{bmatrix} = \begin{bmatrix}
\end{bmatrix}^{T}
\]

Selection of species profile  
Substitution of species profile

Folding species profile  
Unfolding species profile

Trilinearity Constraint  
Unique Solutions!

Loadings give the relative amounts!

PCA, SVD  
1st score gives the common shape

R. Tauler, I. Marqués, and E. Casassas.  
Journal of Chemometrics, 1998; 12, 55-75
Quality assessment of MCR-ALS results

MCR solutions are not unique

Evaluation of rotation ambiguities


\[
D = C S^T + E = D^* + E
\]

\[
S_{\text{new}}^T = T S^T
\]

\[
C_{\text{new}} = C T^{-1}
\]

\[
D^* = C S^T = C T^{-1} T S^T = C_{\text{new}} S_{\text{new}}^T
\]

Rotation matrix \( T \) is not unique. It depends on the constraints.

\( T_{\text{max}} \) and \( T_{\text{min}} \) may be found by a non-linear constrained optimization?
Extensión to ‘multiway’ data: 4 chromatographic runs of 4 coeluting components Trilinear data

\[ d_{ijk} = \sum_{n=1}^{N} c_{in} s_{jn} t_{kn} + e_{ijk} \]

a) Matrix augmentation, non-negativity and spectra normalization constraints

b) Matrix augmentation, non-negativity, spectra normalization and selectivity constraints

c) Matrix augmentation, non-negativity, spectra normalization and trilinearity constraints
Quality assessment of MCR-ALS results.
Error propagation and Confidence intervals

Resampling Methods

Theoretical Data

Experimental Data

Montecarlo Simulation

Noise Addition

Jackknife

Noise 1%

<table>
<thead>
<tr>
<th></th>
<th>pK₁</th>
<th>std.dev</th>
<th>pK₂</th>
<th>std.dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noise added</td>
<td>Value</td>
<td>std.dev</td>
<td>Value</td>
<td>std.dev</td>
</tr>
<tr>
<td>0 %</td>
<td>3.6539</td>
<td>2e-14</td>
<td>4.9238</td>
<td>2e-14</td>
</tr>
<tr>
<td>0.1 %</td>
<td>3.6540</td>
<td>6e-4</td>
<td>4.9226</td>
<td>0.0022</td>
</tr>
<tr>
<td>1 %</td>
<td>3.6592</td>
<td>0.0061</td>
<td>4.9134</td>
<td>0.0264</td>
</tr>
<tr>
<td>2 %</td>
<td>3.6656</td>
<td>0.0101</td>
<td>4.9100</td>
<td>0.0409</td>
</tr>
<tr>
<td>5 %</td>
<td>4.0754</td>
<td>0.4873</td>
<td>5.3308</td>
<td>1.1217</td>
</tr>
</tbody>
</table>
Noise Addition Simulations
Spectra profiles:
Mean, max and min profiles
Confidence range profiles

0% noise

0.1% noise

1% noise

2% noise

5% noise
Until now MCR-ALS input has 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 new graphical user-friendly interface for MCR-ALS

Joaquim Jaumot, Raimundo Gargallo, Anna de Juan and Roma 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 a single experiment

Melting experiment of an oligonucleotide (2 components) monitored by UV-VIS spectroscopy
### Selection of ALS constraints

<table>
<thead>
<tr>
<th>No-negativity</th>
<th>Implementation for conc</th>
<th>Implementation for spec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes?</td>
<td>fnnis</td>
<td>fnnis</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unimodality</th>
<th>Implementation of the unimodality constraint</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes?</td>
<td>average</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Closure</th>
<th>Nr. of closure constraints to be included?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes?</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Equality constraints in conc profiles</th>
<th>Equality constraints in spectra profiles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes?</td>
<td>Yes?</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Optimization parameters</th>
<th>Nr. of iterations</th>
<th>Convergence criterion</th>
<th>Graphical output</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>500</td>
<td>0.1</td>
<td>✔</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Output</th>
<th>Concentration</th>
<th>Std. dev.</th>
<th>Area opt</th>
<th>Optimize</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>eopt</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Spectra</th>
<th>Residuals</th>
<th>Ratio opt</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>sopt</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Example 2. Analysis of multiple experiments. Analysis of 4 HPLC-DAD runs each of them containing four compounds.
Plots are optimum in the iteration Nr. 12

Convergence is achieved!!!

Std. dev of residuals vs. exp. data = 0.0029793
Fitting error (lack of fit, lof) in % (PCA) ~0.43633
Fitting error (lack of fit, lof) in % (exp) ~1.9925
Percent of variance explained ($r^2$) at the optimum is ~99.9603
Spectroscopic Imaging Data
Coupling microscopy and spectroscopy

Number of pixels (x x y)

D Data set

Scanned surface

Pixel

Chemical measurement

\( \lambda \)

\( y \)

\( x \)
Example: Industrial spectroscopic images

Oil-in-water emulsion
(60 × 60 pixels)
Spectroscopic technique
RAMAN (229 wavenumbers)
Compounds in the emulsion overlap.
The interphase is complex.

Oil-in-water emulsion monitored at different depths
(24 x 23 x 10 pixels)

Multivariate Curve Resolution
(spectroscopic image systems modelling)

\[ D = C \]

\[ C = ((m \times n) \times s) \]

\[ D \] unfolded/matrixed

\[ C \] reshape as sampling surface

\[ S \] pure spectrum

\[ S = (s \times p) \]

\[ D = ((m \times n) \times p) \]

\[ D \] distribution map

\[ (m \times n) \]
Resolution of augmented image data sets

MCR-ALS Resolution of spectroscopic images
(only non-negativity constraints)

Distribution maps

Spectra

A

B

C

D
Study of the cis platination reaction between:

Oligopeptide methionine-guanine conjugate
Phac-Met-linker-p^5dG (Phac = phenylacyl)

+ 

cisplatin
[\textsuperscript{15}N]-cis-dichlorodiammineplatinum(II)

2D-NMR reaction monitoring

\[ [^{1}H,^{15}N]-HSQC \text{ NMR correlated spectroscopy} \]

\[ \text{Experimental conditions of the reaction} \]

\[ \text{Met \ldots \ldots \ldots dG} \]
\[ + \]
\[ \text{cisplatin} \]

- reaction time: hours (slow reaction)
- at several times one 2D-NMR spectrum
- 23 2D NMR correlated spectra were measured

\[ \text{Study of this reaction with time} \]
\[ 48 \text{ h} \]

\[ \text{D (23, 156, 751)} \]

\[ \text{15N labelled cisplatin} \]
\[ 751 \text{ 15N chemical shifts} \]
\[ 156 \text{ 1H chemical shifts} \]
Data structure: multiple 2D data matrices at different reaction stages

\[
\begin{align*}
D_{1} & : 156 \delta^{1}H \\
D_{2} & : 751 \delta^{15}N \\
D_{23} & : 156 \delta^{1}H \times 751 \delta^{15}N
\end{align*}
\]
MCR-ALS applied to multiple 2D NMR data matrices

\[
\begin{array}{c|c|c|c|c|c}
D_1 & S_1^T & S_2^T & S_3^T & S_4^T \\
D_2 & S_1 & S_2 & S_3 & S_4 \\
D_{23} & C & 156 \delta^1H & 751 \delta^{15}N & 156 \delta^1H & 751 \delta^{15}N \\
\end{array}
\]

Refolding

Kinetic information:
distribution diagram and 2D “pure” NMR spectra

156 \delta^1H \times 751 \delta^{15}N
D “tube-wise”
Distribution diagram

- NH3 Trans Met Monofunctional
- Final Product
- NH3 Trans dG Monofunctional
- NH3 Trans Metal Chelate

Relative Concentration vs Reaction time (hours)
Do the presence of Ca(II) affecting the protein folding mechanism?

\[ \alpha\text{-lactalbumin} \] (Ca\(^{2+}\) presente)

\[ \alpha\text{-apolactalbumin} \] (Ca\(^{2+}\) ausente)
The protein folding pathway (Far- and near-UV CD)

**α-lactalbumin**

<table>
<thead>
<tr>
<th>D1</th>
<th>D2</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td></td>
</tr>
</tbody>
</table>

**α-apolactalbumin**

<table>
<thead>
<tr>
<th>S1T</th>
<th>S2T</th>
</tr>
</thead>
<tbody>
<tr>
<td>ST</td>
<td></td>
</tr>
</tbody>
</table>

**CD near UV**

<table>
<thead>
<tr>
<th>D</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>250</td>
<td>300</td>
</tr>
</tbody>
</table>

**CD far UV**

<table>
<thead>
<tr>
<th>D</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>200</td>
<td>250</td>
</tr>
</tbody>
</table>
The protein contributions are successfully modelled in the presence of the evolving solvent background.
T-induced transitions of β-lactoglobulin

Protein process description

- The proposed mechanism is confirmed.
- Only the combined use of MIR/NIR defines all the protein conformations.

*S. Navea, A. de Juan and R. Tauler* *Analytical Chemistry*, 2003, 75, 5592-5601
MCR-ALS applied to DNA microarray data

DNA microarray technology has made possible to monitor gene expression levels for thousands of genes in a single experiment.

Information about the existence of patterns and relationships between samples (cell lines) and variables (genes) can be obtained.

Because of the huge amount of data generated in a single experiment, data compression and data analysis methods are needed to extract and understand the information contained in the data.
MCR-ALS applied to DNA microarray data

Experimental data

Initial Estimation
K-means Centroids

ALS

Data Matrix

Samples profiles

Gene profiles

ST

information about the cancer samples (cell lines)

information about the gene expression (variables)
EXTRACTING BIOMEDICAL INFORMATION FROM GENE EXPRESSION MICROARRAY DATA BY MULTIVARIATE CURVE RESOLUTION
J.Jaumot, R.Tauler and R.Gargallo (work in progress)

MCR-ALS results shows that each group is characterized by:

<table>
<thead>
<tr>
<th>MCR Component</th>
<th>Type of cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CNS, OV, RE, LC</td>
</tr>
<tr>
<td>2</td>
<td>ME</td>
</tr>
<tr>
<td>3</td>
<td>LE</td>
</tr>
<tr>
<td>4</td>
<td>Not well defined</td>
</tr>
<tr>
<td>5</td>
<td>CO</td>
</tr>
</tbody>
</table>

LE leukemia, ME melanoma
CO colon, other carcinomas

C (cell lines)

S^T (genes)
QUANTITATION USING MCR-ALS

(A) $m/z$

(MCR-ALS) $D = C + E$

(B) $D_{aug} = C_{aug} + E_{aug}$
Reconstructed TIC MS chromatograms

Standard mixture of 13 biocides at 20 ppm

Aznalcollar sediment sample spiked with 13 biocides at 10 ppm

In-WWTP water sample from La LLagosta sample spiked with 13 biocides at 10 ppm
MULTIVARIATE CURVE RESOLUTION

\[ t_R \quad \lambda \quad m/z + \text{wavelets} \]

\[ \text{DAD} \quad \text{MS} \]

\[ \text{DAD} \quad \text{MS} \]

\[ \text{DAD and MS Spectra} \]

\[ \text{Elution profiles} \]

\[ \text{alachlor} \quad \text{chlopyrifos-oxon} \quad \text{terbutryn} \]

\[ \text{solvent gradient} \]

\[ \text{SCAN mode} \]
MCR-ALS resolved elution profiles for standard mixture samples of 20, 10, 5, 1, 0.5 and 0.1 ppm together with standard pure samples of alachlor, chlorpyrifos and terbutryn

Determinació directa d’analits a partir de mesures espectrofluorimètriques en matrius naturals complexes sense necesitat de separació prèvia per métodes cromatogràfics

Determinació de trifenilestany en aigua de mar mitjançant fluorescència i resolució multivariant

Analytica Chimica Acta, 2000, 409, 237-245
MCR-ALS resolution of [U;S;R;B] augmented matrix

a) 3-D plots of the EEM fluorescence of the unknown sample U, standard S, flavonol reagent R and sea-water background B;

b) emission spectra for the unknown sea-water sample; c) emission species spectra for the standard;
d) emission species spectra for flavonol reagent; e) emission species spectra for sea-water background;
f) excitation spectra
Resolución y quantificación por MCR-ALS de datos EEM

Areas de los espectros de emisión resueltos por MCR-ALS respecto a las concentraciones de TPhT

Estratègias de Calibración
i. using pure standards
ii. using sea-water standards
iii. using the standard addition method

Comparación valores verdaderos y calculados por MCR-ALS en muestras de aguas de mar

\[ c_U = \frac{\text{Area}(y_U)}{\text{Area}(y_S)} \cdot c_S \]

errores de predicción siempre por debajo del 13%!

Standard addition calibration graph in a sea-water analyte determination (sea-water sample U4)

The Analyst, 2000, 125, 2038-43
Environmental data tables (two-way data)

\[ X(I,J) \]

Data table or data matrix

- I samples
- J variables

Plot of samples (rows)

Plot of variables (columns)

Conc. of chemicals
Physical Properties
Biological properties
Other .....
Environmental source resolution and apportioment

**Bilinearity!**

\[ x_{ij} = \sum_{n=1}^{N} g_{nj} f_{in} + e_{ij} \]

- \( x_{ij} \): concentration of chemical contaminant \( j \) in sample \( i \)
- \( f_{in} \): contribution of source \( n \) in sample \( j \)
- \( g_{nj} \): contribution of compound \( j \) in source \( n \)

- **G**:
  - Identification of contamination sources (composition)
  - Distribution of contamination sources

- **E**:
  - 22 samples
  - Identification of contamination sources

- **X**:
  - Concentration of 96 organic compounds

- **NR**:
  - Contribution of compound \( j \) in source \( n \)
Projecte ACA
E. Peré-Trepat, M. Terrado, R. Tauler,
Contaminación general
AQUATERRA
Sub-Project BASIN

Workpackage: R2 – EBRO river basin
Integration Monitoring-Chemometrics-GIS
M.Terrado, A.Navarro, S.Lacorte and R.Tauler

WP Leader: D. BARCELO
IIQAB-CSIC, Barcelona (E)
Partners:
ACA, Barcelona (E)
AGBAR, Barcelona (E)
AGH, Krasow (PL)
SURFACE WATER

AQUATERRA (Surveillance monitoring)
- R0: Ebro in Reinosa (Cantabria)
- R1: Ebro in Miranda de Ebro (Burgos)
- T2: Zadorra in Audinaka (Álava)
- T3: Zadorra in Villodas (Álava)
- R4: Ebro in Haro (La Rioja)
- T5: Najerilla in Najera (La Rioja)
- R6: Ebro in Logroño (La Rioja)
- T7: Ega in Estella (Navarra)
- R7: Ebro in Tudela (Navarra)
- T8: Araquil in Alcàsser (Navarra)
- T9: Arga in Puente la Reina (Navarra)
- T10: Jalón in Grisén (Zaragoza)
- T11: Huerva in Zaragoza (Zaragoza)
- T12: Gállego in Caldearenas (Huesca)
- T13: Gállego in San Mateo de Gállego (Zaragoza)
- R14: Ebro in Presa de Pina (Zaragoza)
- R15: Ebro in Sástago (Zaragoza)
- T16: Segre in Torres de Segre (Lleida)
- R17: Ebro in Flix (Tarragona)
- R18: Ebro in Tortosa (Tarragona)
- R19: Ebro in Amposta (Tarragona)
- R20: Ebro in Delta de l’Ebro (Tarragona)
- GAR1: Gállego in Villanueva de Gállego (Zaragoza)

24 sample points

RED ICA → CHE (Red Integral de Control de Aguas)

RED ICA → CHE (Red de control de sustancias peligrosas)
- SP-1: Gállego en Jabarrella
- SP-2: Ebro en Presa Pina
- SP-3: Ebro en Ascó
- SP-4: Segre en Torres de Segre
- SP-5: Cinca en Monzón (aguas abajo)
- SP-6: Arga en Puente la Reina
- SP-7: Ebro en Miranda de Ebro
- SP-8: Zadorra en Vitoria Trespuentes
- SP-9: Ebro en Tortosa
- SP-10: Araquil en Alcàsser-Urdiain
- SP-11: Ebro en Conchas de Haro
- SP-12: Ebro en Logroño (aguas abajo)-Varea
- SP-13: Ega en Arinzano
- SP-14: Gállego en Villanueva
- SP-15: Huerva en Fuente de la Junquera
- SP-16: Jalón en Grisén
- SP-17: Najerilla en Nájera (aguas abajo)
- SP-18: Zadorra en Salvatierra

210 sample points

18 sample points
SURFACE WATER → HISTORICAL DATA

1. DANGEROUS SUBSTANCES CONTROL NETWORK
   (Red de Control de Sustancias Peligrosas)
   18 sampling points; 3 compartments comparison

   WATER (2002-04)  SEDIMENTS (2002-03)  BIOTA (2002-03)

2. ICA NET OF SURFACE WATER CONTROL
   (Red Integrada de Control de la Calidad de las Aguas Superficiales)
   Selection of some points from this net common with the AQUATERRA's surveillance monitoring:
   1 compartment

   WATER (2002-04)

Physical parameters
- Flow
- Water temperature
- Air temperature
- Conductivity at 20°C
- Aspect

Chemical parameters
- pH
- Dissolved oxygen
- Suspended materials
- COD (chemical oxygen demand), BOD5...
- Chlorides, sulphates, phosphates

METALS: As, Cd, Pb, Zn, Cu
ORGANIC COMPOUNDS: pesticides of agricultural origin (HCB, SDDTs, SHCHs)
- SPCBs
- PAHs

Other parameters
- LAND USES → grouping points according to their land use, to obtain common features and differences
- QUALITY INDEXES:
  - ISQA (simplified index for water quality)
  - BIOLOGIC index
SURFACE WATER AND GROUNDWATER DATABASES

CHEMOMETRICS

PCA (Principal Component Analysis)

Other methods (Multivariate data Analysis)

IDENTIFICATION
contamination sources

DISTRIBUTION
contamination sources

Loading plots

Geographical

Temporal

GIS

ENVIRONMENTAL DISCUSSION

- Key zone delimitation according to the pollution degree.
- Relation between land use (agricultural, industrial, urban and protected areas) and distribution of the contamination sources
- Representation of quality indexes
- Comparison of the contribution of the different contamination sources in each environmental compartment (water, sediments and biota)
Acknowledgements

• **Chemometrics Group UB**
  – **Staff:** Anna de Juan, Javier Saurina, Raimundo Gargallo (RyC)
  – **PhD:** Susana Navea, Joaquim Jaumot, Emma Peré-Trepat
  – **Master and DEA:** Gloria Muñoz, Silvia Mas

• **Environmental Chemometrics Group IIQAB-CSIC**
  – **Staff:** Romà Tauler
  – **Post-doc:** Montse Vives (JdC), Mónica Felipe (I3P)
  – **Master and DEA** Marta Terrado, Xavier Puig

• Elisabeth Teixido (ACA), Silvia Termes (LAG)