Estimation of error propagation and prediction intervals in Multivariate Curve Resolution Alternating Least Squares using resampling methods

Joaquim Jaumot, Raimundo Gargallo and Romà Tauler*
Department of Analytical Chemistry
University of Barcelona
Outline:

• Introduction
• Rotational ambiguities and feasible bands
• Error propagation and resampling methods
• Results
• Conclusions
Multivariate (Soft) Self Modeling Curve Resolution

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

Bilinearity!
Multivariate (Soft) Self Modeling Curve Resolution

- Multivariate Curve Resolution (MCR) methods have been shown to be powerful self-soft-modeling tools able to investigate complex chemical systems with a minimum number of assumptions.

- Alternating Least Squares (ALS) has become a popular method for Multivariate Curve Resolution (MCR) due to its flexibility in constraint implementation during the optimization of resolved profiles.
Multivariate (Soft) Self Modeling Curve Resolution

- **What are the reliability of MCR-ALS estimations?**
  - Do the MCR-ALS solutions have rotational and scale freedom?
  - Are they unique solutions or exist instead a band of feasible solutions?
  - How errors and noise are propagated from experimental data to ALS estimations?
Goals of this study

• Find the reliability of ALS resolved profiles in multivariate curve resolution.

• Estimate prediction error intervals for ALS profiles

• Estimate prediction error intervals for parameters calculated from MCR-ALS resolved profiles

• Investigate the interaction between propagation of errors and rotational ambiguities (noise effects on rotational ambiguities and constraints).
Outline:

• Introduction
• **Rotational ambiguities and calculation of feasible bands**
• Error propagation and resampling methods
• Results
• Conclusions
Lawton and Sylvestre feasible bands

W.H. Lawton and EA Sylvestre, Technometrics 1971, 13, 617-33
Rotational Ambiguities

Factor Analysis (PCA) Data Matrix Decomposition
\[ D = U \ V^T + E \]

‘True’ Data Matrix Decomposition
\[ D = C \ S^T + E \]

\[ D = U \ T \ T^{-1} \ V^T + E = C \ S^T + E \]

\[ C = U \ T; \quad S^T = T^{-1} \ V^T \]

How to find the rotation matrix \( T \)?

Matrix decomposition is not unique!
\( T(N,N) \) is any non-singular matrix
There is rotational freedom for \( T \)
Rotational Ambiguities

Because of rotational ambiguities instead of unique solutions, a set of *feasible solutions* are obtained.

Feasible solutions are different solutions that fit equally well the data under a set of constraints.

For a particular system under a set of constraints, feasible solutions are defined from a set of possible $T$ values.
Rotational Ambiguities

- $T$ values define the band of feasible solutions or feasible bands
  - How to define the boundaries of these feasible bands?
  - How to represent graphically these boundaries?
Is it possible to define band boundaries ($T_{\text{max}}$ and $T_{\text{min}}$)?

How to calculate $T_{\text{max}}$ and $T_{\text{min}}$?
How to define and find the band boundaries?

• What are the $T$ values giving the *maximum/outer and minimum/inner boundaries* of the feasible bands under a set of constraints?

\[
D^* = C_{inic} S_{inic}^T =
\]

\[
= C_{inic} T_{min} T_{min}^{-1} S_{inic}^T = C_{min} S_{min}^T =
\]

\[
= C_{inic} T_{max} T_{max}^{-1} S_{inic}^T = C_{max} S_{max}^T
\]

where: $D(NR,NC)$, $C(NR,N)$, $S^T(N,NC)$, $T(N,N)$

How to define and evaluate $T_{max}$ and $T_{min}$?
Evaluation of boundaries of feasible bands: Previous studies

• W.H.Lawton and E.A.Sylvestre, Technometrics, 1971, 13, 617-633
• P. Gemperline (Analytical Chemistry, 1999, 71, 5398-5404)
• R.Tauler (J.of Chemometrics 2001, 15, 627-46)
Definition of band boundaries

The whole measured signal is:
\[ D = \sum D_i = \sum c_i s_i^T \]

The contribution of each species to the whole signal is:
\[ D_i = c_i s_i^T \]

Solving the Optimization Problem:

max/outer boundary: Find \( T_{\text{max}} \) that makes \( c_i s_i^T \) maximum

min/inner boundary: Find \( T_{\text{min}} \) that makes \( c_i s_i^T \) minimum
Constrained Non-Linear Optimization Problem (NCP)

Find $T$ which makes:

\[
\begin{aligned}
\min/\max \ f(T) & \quad \text{subject to} \quad g_e(T) = 0 \\
T & \quad \text{and to} \quad g_i(T) \leq 0
\end{aligned}
\]

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

Matlab Optimization Toolbox \texttt{fmincon} function
1) What are the variables of the problem?

\( T \) (rotation matrix),
\[ D = C \, T \, T^{-1} \, S^T \]

2) What is the objective function \( f(T) \) to be optimized?

For each species \( i = 1, \ldots, n_s \)

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

This gives the relative signal contribution of species \( i \) respect the global measured signal!

\( f(T) \) is scalar value between 0 and 1!
3) What are the constraints $g(T)$?

The following constraints may be 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 \( g_{\text{unim}} \)
- trilinearity (three-way data) \( g_{\text{tril}} \)

Are they equality or inequality constraints?
4) What are the initial estimates of C, $S^T$?

- Initial estimates of C and $S^T$ are obtained by MCR-ALS
- Initial estimates are feasible solutions fulfilling the constraints of the system (*non-negativity, unimodality, closure, selectivity, local rank,...*)

5) What are the initial values of T?

- NCP depends on initial estimates of T! (local minima, convergence, speed …)

$$T_{ini} = \text{eye}(N) = \begin{pmatrix}
1 & 0 & \ldots & 0 \\
0 & 1 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & 1
\end{pmatrix}$$
Optimization algorithm

• R. Tauler (J. of Chemometrics 2001, 15, 627-46)

Initial estimations of $\mathbf{C}_{\text{ALS}}$ and $\mathbf{S}_{\text{ALS}}$
profiles are obtained by MCR-ALS
$\mathbf{T} = \text{eye}(\text{number of species})$

For each species define objective function
$f(\mathbf{T}) = \text{norm}(\mathbf{c}(\mathbf{T})\mathbf{s}(\mathbf{T})) = \text{norm}(\mathbf{c}_{\text{ALS}} \mathbf{T} \mathbf{s}_{\text{ALS}} / \mathbf{T})$

Select constraints $g(\mathbf{T})$:
equality $g_e$: normalization/closure, known values,
inequality $g_i$: non-negartivity, selectivity, unimodality, trilinearity,

Find $\mathbf{T}_{\text{min}}$ which gives a minimum of $f(\mathbf{T})$
under constraints $g_i(\mathbf{T})<0$, $g_e(\mathbf{T})=0$
Built minimum band
$\mathbf{c}_{\text{min}} = \mathbf{c}_{\text{ALS}} / \mathbf{T}_{\text{min}}$
$\mathbf{s}_{\text{min}} = \mathbf{s}_{\text{ALS}} / \mathbf{T}_{\text{min}}$

Find $\mathbf{T}_{\text{max}}$ which gives a maximum of $f(\mathbf{T})$
under constraints $g_i(\mathbf{T})<0$, $g_e(\mathbf{T})=0$
Built maximum band
$\mathbf{c}_{\text{max}} = \mathbf{c}_{\text{ALS}} / \mathbf{T}_{\text{max}}$
$\mathbf{s}_{\text{max}} = \mathbf{s}_{\text{ALS}} / \mathbf{T}_{\text{max}}$
Experimental data system under study

Acid-base spectrophotometric titration of the double stranded heteropolynucleotide polyinosinic-polycytidylic acid. Spectral region between 240-320 nm and pH region between pH 2 and pH 9
Application of MCR-ALS to the experimental data matrix D

Applied constraints in ALS were:

a) non-negative spectra
b) non-negative concentrations
c) closure in concentrations

Initial estimates were obtained from purest variables

• This system has selectivity! local rank resolution conditions!
• Initial estimates from pure variable detection methods provide good initial estimates that produce solutions close to the true profiles
Parameter estimation

Mass-action law is only assumed at the site level and not for the whole polynucleotide molecule

<table>
<thead>
<tr>
<th>Evaluation of constants from intersection profiles</th>
<th>Proposed species:</th>
</tr>
</thead>
<tbody>
<tr>
<td>pK$_1$</td>
<td>3.6660</td>
</tr>
<tr>
<td>pK$_2$</td>
<td>4.9244</td>
</tr>
</tbody>
</table>
Estimation of band boundaries
(max/min contribution of each species)
of feasible solutions

Large Rotational ambiguities were present when constraints applied were only closure non-negativity!!!
Estimation of band boundaries
(max/min contribution of each species)
of feasible solutions

Rotational ambiguities nearly disappear when selectivity constraint was applied!!!
Outline:

• Introduction
• Rotational ambiguities and feasible bands
• Error propagation and resampling methods
• Results
• Conclusions
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
Resampling Methods

Theoretical Data

Montecarlo Simulation

Experimental Data

Noise Addition

Jackknife
Building theoretical data

Experimental Data, $D_{\text{exp}}$

MCR-ALS

C

$S^T$

Theoretical Data, $D$

Experimental error $E$
Montecarlo Simulations

\[ M_{0.1} = D + N_{0.1} \]
\[ M_1 = D + N_1 \]
\[ M_2 = D + N_2 \]
\[ M_5 = D + N_5 \]

Theoretical Data \( D \) 
New Data Matrix 
Random Error \( N_{0.1}, N_1, N_2 \) and \( N_5 \) 
250 times each noise level! 
1000 simulations!

MATLAB function \texttt{randomn} with zero mean and relative sd 0.1%, 1%, 2% and 5% of maximum signal in \( D \)
Noise Addition Simulations

\[ D_{0.1} = D_{\text{exp}} + N_{0.1} \]
\[ D_{1} = D_{\text{exp}} + N_{1} \]
\[ D_{2} = D_{\text{exp}} + N_{2} \]
\[ D_{5} = D_{\text{exp}} + N_{5} \]

MATLAB function *randomn* with zero mean and relative sd 0.1%, 1%, 2% and 5% of maximum signal in \( D \)

250 times each noise level!
1000 simulations!
<table>
<thead>
<tr>
<th>N₀.₁ N₁ N₂ and N₅</th>
<th>Experimental Data D_{\text{exp}} (36,81)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+ Random Error</td>
</tr>
<tr>
<td></td>
<td>= New Data Matrix D_{\text{noise}} (36,81)</td>
</tr>
</tbody>
</table>

| J₁               | Reduced Matrix (1,10,19,28) (32,81)   |
| J₂               | Reduced Matrix (2,11,20,29) (32,81)   |
| J₃               | Reduced Matrix (3,12,21,30) (32,81)   |
| J₄               | Reduced Matrix (4,13,22,31) (32,81)   |
| J₅               | Reduced Matrix (5,14,23,32) (32,81)   |
| J₆               | Reduced Matrix (6,15,24,33) (32,81)   |
| J₇               | Reduced Matrix (7,16,25,34) (32,81)   |
| J₈               | Reduced Matrix (8,17,26,35) (32,81)   |
| J₉               | Reduced Matrix (9,18,27,36) (32,81)   |
Jackknife Simulations

Jack Knife Reduced Data Matrix $J_N$

N = 1,..., 9

MCR-ALS

Concentration profiles $C$

Pure Spectra $S_T$
Outline:

• Introduction
• Rotational ambiguities and feasible bands
• Error propagation and resampling methods
• Results
• Conclusions
Presentation of Results

1. Calculation of species profiles error bands: Mean profile, maximum and minimum profiles, standard deviation profiles and confidence range profiles

2. pKa (parameter) error estimations

3. Rotational ambiguity effects on error estimates from resampling methods. Calculation of boundaries of feasible bands from mean species profiles error bands
Monte Carlo Simulations
Concentration profiles:
Mean max and min profiles
Confidence range profiles

Noise 0%
Noise 0.1%
Noise 1%
Noise 2%
Noise 5%
Monte Carlo Simulations
Spectra profiles:
Mean max and min profiles
Confidence range profiles
## Monte Carlo Simulations
### pKα error estimations

<table>
<thead>
<tr>
<th>Noise added</th>
<th>pK₁ Value</th>
<th>pK₁ Std. dev</th>
<th>pK₂ Value</th>
<th>pK₂ Std. dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 %</td>
<td>3.6660</td>
<td>4e-15</td>
<td>4.9244</td>
<td>9e-15</td>
</tr>
<tr>
<td>0.1 %</td>
<td>3.6662</td>
<td>6e-4</td>
<td>4.9243</td>
<td>0.0012</td>
</tr>
<tr>
<td>1 %</td>
<td>3.6696</td>
<td>0.0065</td>
<td>4.9262</td>
<td>0.0128</td>
</tr>
<tr>
<td>2 %</td>
<td>3.6761</td>
<td>0.0127</td>
<td>4.9173</td>
<td>0.0245</td>
</tr>
<tr>
<td>5 %</td>
<td>3.9762</td>
<td>0.4349</td>
<td>5.0745</td>
<td>0.7595</td>
</tr>
</tbody>
</table>
Calculation of band boundaries from mean species profiles error bands (under non-negativity and closure constraints)
Calculation of band boundaries from mean profile error bands (under non-negativity, closure and selectivity constraints)

Noise 0.1%

Noise 2%

Noise 1%

Noise 5%
Noise Addition Simulations
Concentration profiles:
Mean max and min profiles
Confidence range profiles
Noise Addition Simulations

Spectra profiles:
Mean, max and min profiles
Confidence range profiles
## Noise Addition Simulations

**pKa error estimations**

<table>
<thead>
<tr>
<th>Noise added</th>
<th>pK₁ Value</th>
<th>Std. dev</th>
<th>pK₂ Value</th>
<th>Std. dev</th>
</tr>
</thead>
<tbody>
<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>
Calculation of band boundaries from mean profile error bands (under non-negativity and closure constraints) at 1% error noise addition
Jackknife Simulations at 1% noise; Concentration profiles: Mean max and min profiles and confidence range profiles.
Jackknife Simulations at 1% noise; spectra profiles:
Mean max and min profiles and confidence range profiles
### Jackknife Simulations

pKa error estimations at 1% noise level

<table>
<thead>
<tr>
<th>Nº exp</th>
<th>pK$_1$</th>
<th>pK$_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.6629 ± 0.0066</td>
<td>4.9135 ± 0.0277</td>
</tr>
<tr>
<td>2</td>
<td>3.6601 ± 0.0074</td>
<td>4.8989 ± 0.0221</td>
</tr>
<tr>
<td>3</td>
<td>3.6590 ± 0.0059</td>
<td>4.9122 ± 0.0261</td>
</tr>
<tr>
<td>4</td>
<td>3.6580 ± 0.0056</td>
<td>4.9221 ± 0.0189</td>
</tr>
<tr>
<td>5</td>
<td>3.6333 ± 0.0130</td>
<td>4.9018 ± 0.0236</td>
</tr>
<tr>
<td>6</td>
<td>3.6882 ± 0.0198</td>
<td>4.9144 ± 0.0267</td>
</tr>
<tr>
<td>7</td>
<td>3.6591 ± 0.0064</td>
<td>4.9144 ± 0.0256</td>
</tr>
<tr>
<td>8</td>
<td>3.6592 ± 0.0059</td>
<td>4.9144 ± 0.0253</td>
</tr>
<tr>
<td>9</td>
<td>3.6582 ± 0.0065</td>
<td>4.9233 ± 0.0239</td>
</tr>
</tbody>
</table>
## Parameter Estimation

### Summary of results

<table>
<thead>
<tr>
<th></th>
<th>Real</th>
<th>0.1 %</th>
<th>1 %</th>
<th>2 %</th>
<th>5 %</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>pk1</td>
<td>pk2</td>
<td>pk1</td>
<td>pk2</td>
<td>pk1</td>
</tr>
<tr>
<td><strong>Theoretical Value</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value</td>
<td>3.6660</td>
<td>4.9244</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>MonteCarlo</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simulations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stand. dev.</td>
<td>-</td>
<td>-</td>
<td>0.0006</td>
<td>0.0012</td>
<td>0.0065</td>
</tr>
<tr>
<td><strong>Noise Addition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stand. dev.</td>
<td>-</td>
<td>-</td>
<td>0.0006</td>
<td>0.0022</td>
<td>0.0061</td>
</tr>
<tr>
<td><strong>JackKnife</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stand. dev.</td>
<td>-</td>
<td>-</td>
<td>0.0038</td>
<td>0.0032</td>
<td>0.0086</td>
</tr>
<tr>
<td>Theoretical Value</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Value</td>
<td>3.6660</td>
<td>4.9244</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td><strong>MonteCarlo</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Simulations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stand. dev.</td>
<td>-</td>
<td>-</td>
<td>0.0006</td>
<td>0.0012</td>
<td>0.0065</td>
</tr>
<tr>
<td><strong>Noise Addition</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stand. dev.</td>
<td>-</td>
<td>-</td>
<td>0.0006</td>
<td>0.0022</td>
<td>0.0061</td>
</tr>
<tr>
<td><strong>JackKnife</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Stand. dev.</td>
<td>-</td>
<td>-</td>
<td>0.0038</td>
<td>0.0032</td>
<td>0.0086</td>
</tr>
</tbody>
</table>
Outline:

• Introduction
• Rotational ambiguities and feasible bands
• Error propagation and resampling methods
• Experimental system and simulations
• Results
• Conclusions
Summary

• Different approaches for calculation of error propagation and prediction intervals of estimations have been compared including: Monte Carlo simulations, Noise addition resampling approaches and Jackknife based methods.

• The obtained results allowed a preliminary investigation of the noise effects on MCR-ALS resolved profiles and on parameters from them estimated, and allowed also a preliminary investigation of noise effects on rotational ambiguities.

• The study has been shown for the resolution of a three-component equilibrium system with overlapping concentration and spectra profiles.
Conclusions

- Rotational ambiguity effects on species profiles depend on the structure and constraints of the data system.

- Rotational ambiguities effects at low noise levels in a system with low selectivity are more important than error propagation effects.

- However, at high noise levels ($\geq 5\%$), error propagation effects became larger than rotational ambiguities effects and they are both mixed and undistinguishable.

- Obviously the best is to have a system with enough selectivity (low rotational ambiguities) and with low noise levels (low error propagation).
ELUCIDATION OF THE STRUCTURE OF A PROTEIN FOLDING INTERMEDIATE (MOLTEN GLOBULE STATE) USING MULTIVARIATE CURVE RESOLUTION ALTERNATING LEAST SQUARES (MCR-ALS)
Susana Navea, Anna de Juan and Romà Tauler

MULTIVARIATE CURVE RESOLUTION ALTERNATING LEAST SQUARES ANALYSIS OF THE CONFORMATIONAL EQUILIBRIA OF THE OLIGONUCLEOTIDE d<TGCTCGCT>
Joaquim Jaumot, Núria Escaja, Raimundo Gargallo, Enrique Pedroso and Romà Tauler
HARD AND SOFT MODELLING OF ACID-BASE CHEMICAL EQUILIBRIA OF BIOMOLECULES USING $^1$H-NMR
Joaquim Jaumot, Montserrat Vives, Raimundo Gargallo and Romà Tauler

IDENTIFICATION AND DISTRIBUTION OF MICROCONTAMINANTS SOURCES OF NONIONIC SURFACTANTS, THEIR DEGRADATION PRODUCTS AND LINEAR ALKYLBENZENE SULFONATES IN COASTAL WATERS AND SEDIMENTS IN SPAIN BY MEANS OF CHEMOMETRIC METHODS
Emma Peré-Trepat, Mira Petrovic, Damià Barceló and Romà Tauler

MULTIWAY DATA ANALYSIS OF ENVIRONMENTAL CONTAMINATION SOURCES IN SURFACE NATURAL WATERS OF CATALONIA (SPAIN)
Emma Peré-Trepat, Mónica Flo, Montserrat Muñoz, Manel Vilanova, Josep Caixach, Antoni Ginebreda, Romà Tauler