Methods for the resolution of completely co-eluting components in mass spectrometry data

Katharine M. Mullen
Biophysics Group, Department of Physics and Astronomy
Vrije Universiteit Amsterdam, The Netherlands
kate@few.vu.nl

13 August, 2008

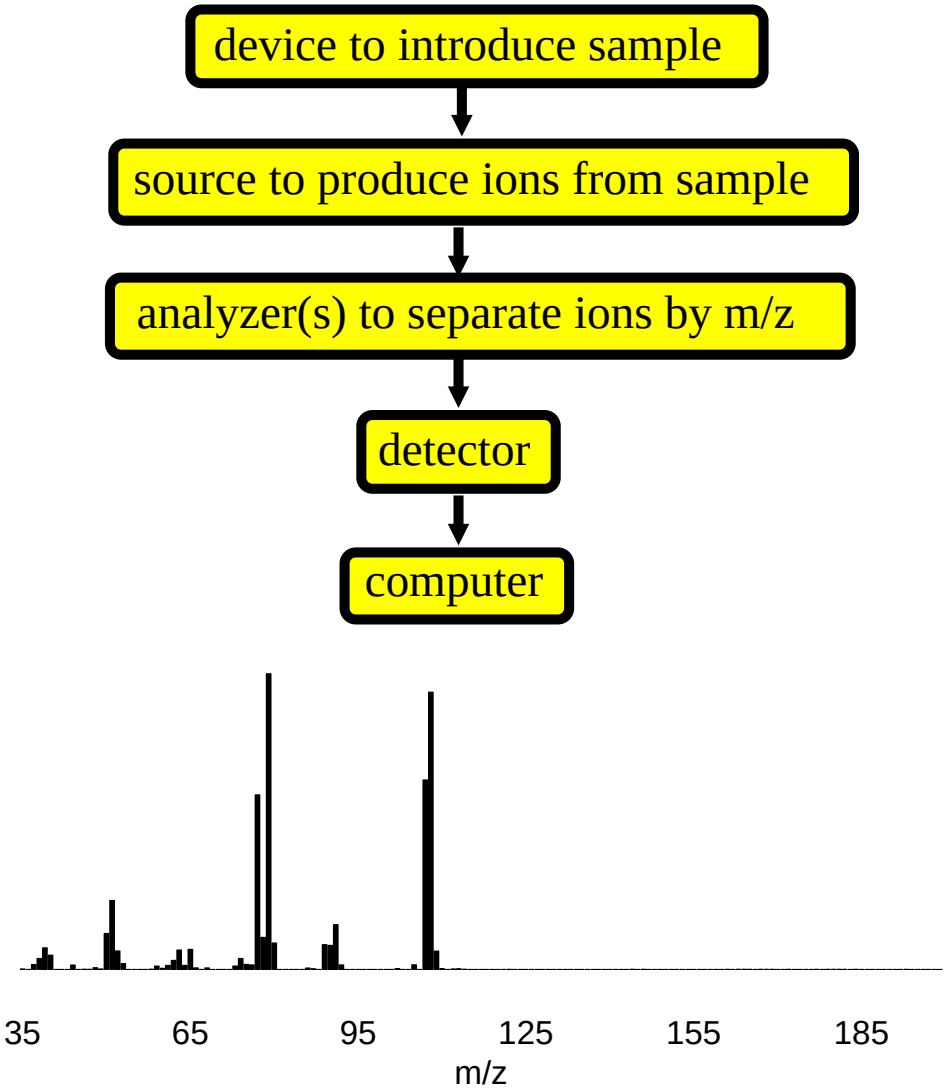
[Talk overview]

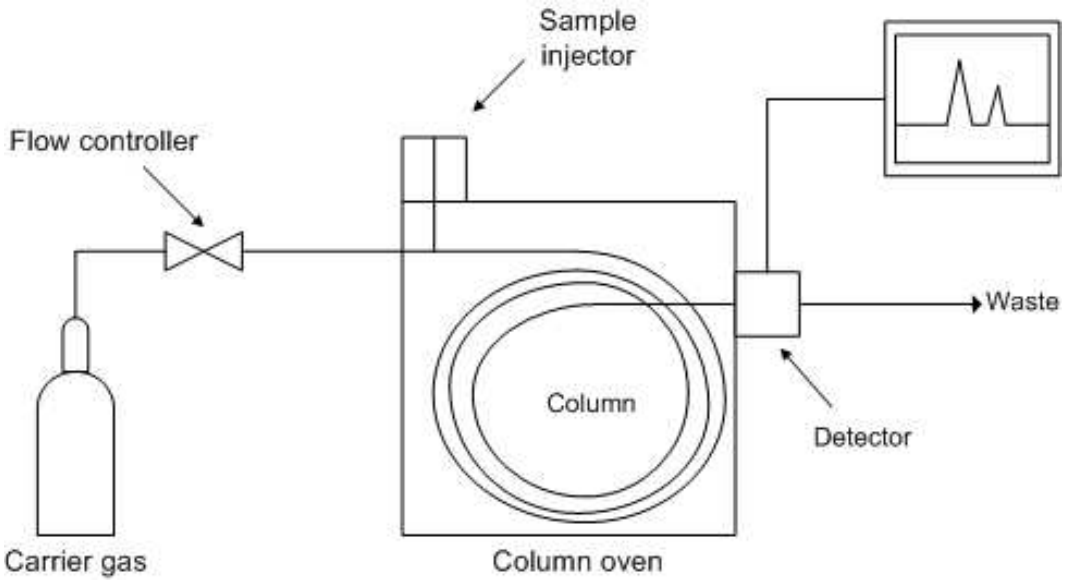
Introduction
- Instrumentation
- Data
- Component resolution problem

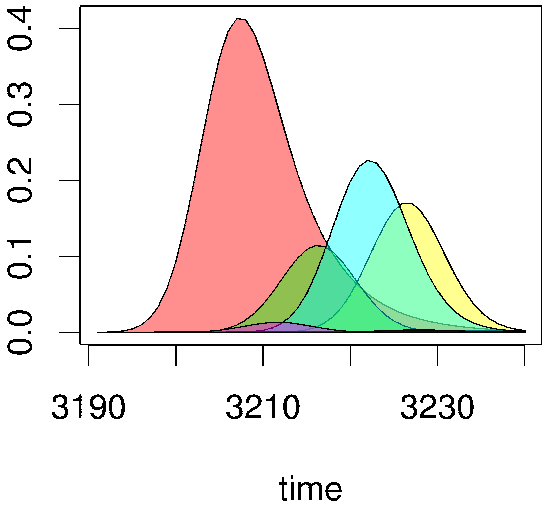
Algorithms for component resolution
- Peak-based
- MCR-ALS
- Global analysis

Resolution of completely co-eluting components
- Initial example
- Robustness of result

Conclusions

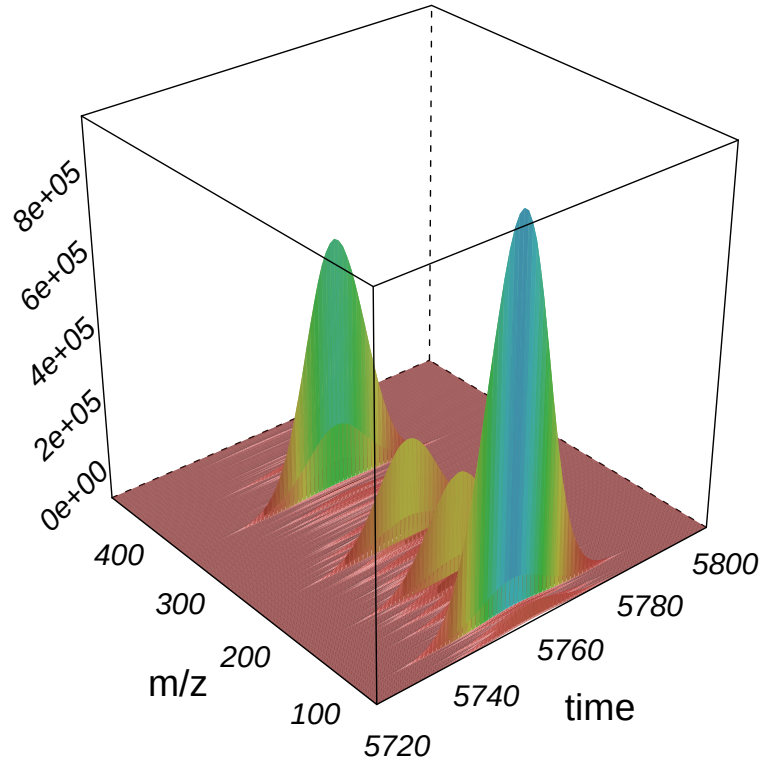

device to introduce sample
source to produce ions from sample
analyzer(s) to separate ions by m/z
detector
computer
35
65
95
125
155
185
m/z


Sample
injector
Flow controller
Waste
Column
Detector
Carrier gas
Column oven


0.0
0.1
0.2
0.3
0.4
3190
3210
3230
time

can use a liquid or gas chromatography column to send sample to the mass spectrometer

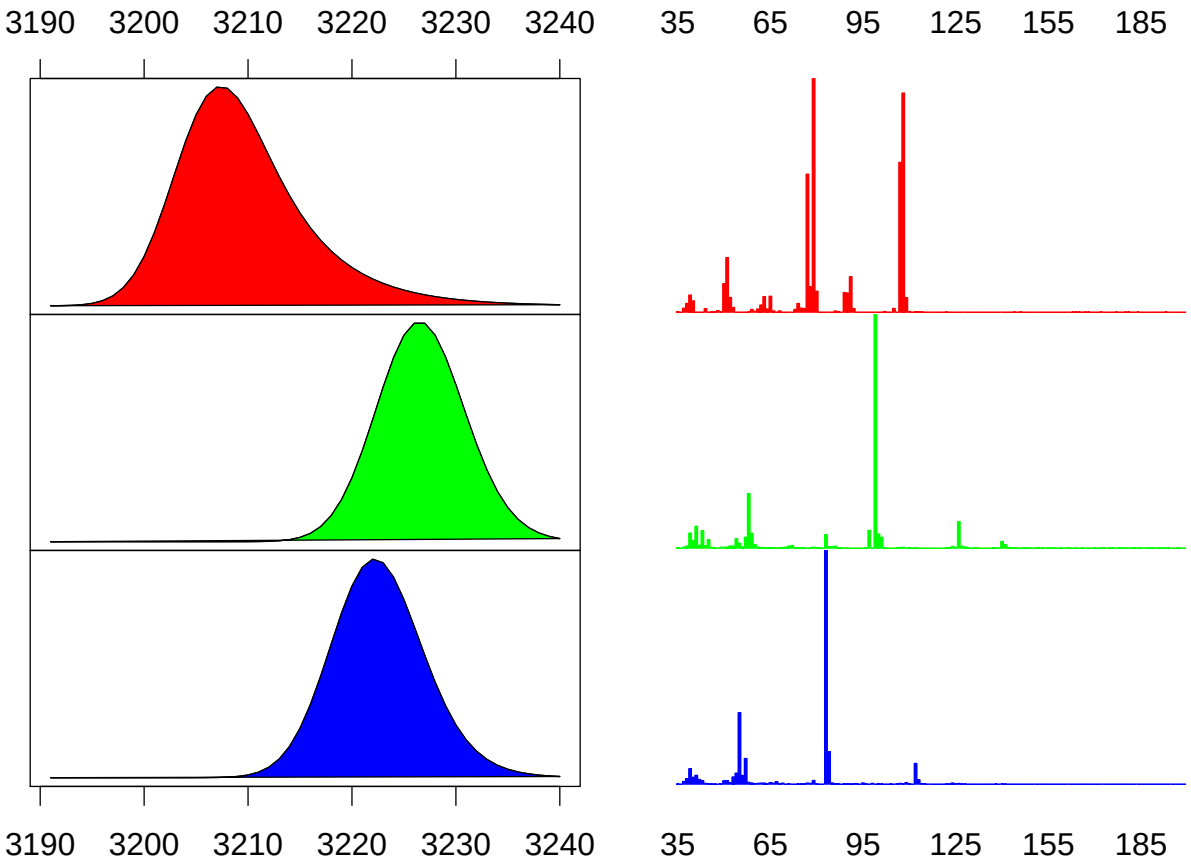
resulting measurement is mass-to-charge ratio resolved with respect to time:



[Component resolution problem]

$D = CS^T$ where D is matrix of data, column i of C is the elution profile of a compound, and column i of S is mass spectrum of that compound

given D, want to estimate the elution profiles C and mass spectra S:



component resolution for multiple datasets $D_1, D_2, \ldots, D_K$:

$$\begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_K \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_K \end{bmatrix} S^T$$

- elution profiles free to vary per-dataset
- mass spectra assumed to be constant between datasets
- integral of an elution profile in different datasets used to quantify relative abundance of component between datasets

Given estimates for mass spectra/retention times, match this information against those of compounds in library (e.g. NIST 05 database) and identify chemical constituents of the sample

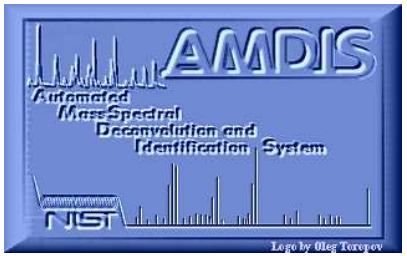
quantify similarity of estimated spectrum to that of spectrum in database via a matching factor function, e.g., the normalized dot product matching factor:

$$MF(u,s) = \frac{u \cdot s}{\|u\| \, \|s\|} = \frac{u_1 s_1 + u_2 s_2 + \ldots}{\sqrt{u_1^2 + u_2^2 + \ldots}\sqrt{s_1^2 + s_2^2 + \ldots}}$$

peak-based methods for component resolution:

- examine the chromatogram at each m/z value for peaks
- examine the total ion chromatogram (TIC) for peaks
- use heuristics to decide which peaks represent components / infer the shape of components
- thereby determine C, from which S can be determined by linear regression from $D = CS^T$

Automated mass spectral deconvolution and identification system (AMDIS) of Stein and collaborators is an example:



[MCR-ALS]

multivariate curve resolution alternating least squares (MCR-ALS) is a self-modeling method: given $D = CS^T$ starts with some starting guess for C then alternates refinement of C and S

- $C = (S^T)^+ D$
- $S^T = C^+ D$

results in estimates for C and S but no parametric model-based description

MCR-ALS does allow solutions for C and S to be subject to constraints:

- non-negativity via R package **nnls**
- unimodality via R package **Iso**
- closure/normalization
- selectivity
- ...

MCR-ALS extends naturally to modeling multiple datasets simultaneously

Implementation of MCR-ALS:

- available as the new R package **ALS**
- all important constraints included
- on CRAN

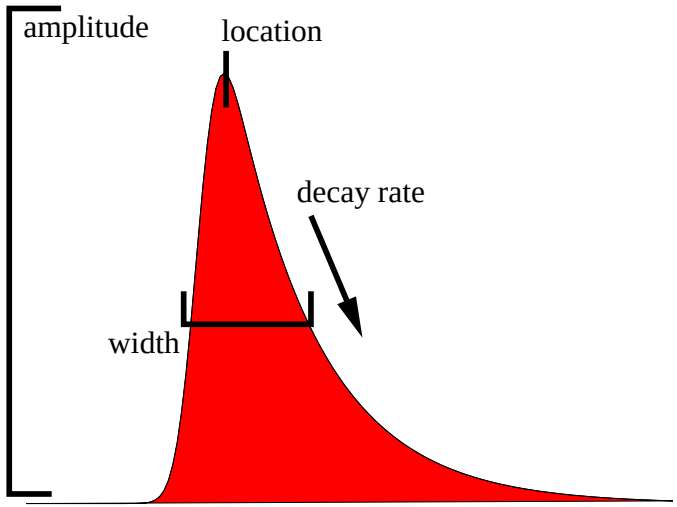
Motivation:

- facilitate a systematic comparison of the component resolution abilities of global analysis and MCR-ALS for data in which underlying elution profiles are completely overlapping

Global analysis:

- elution profiles are described with a parametric model $C(\theta)$
- starting values for θ are improved and estimates for mass spectra S are obtained by solution of a separable nonlinear least squares problem
- mass spectra and amplitude parameters of elution profiles subject to non-negativity constraints
- like MCR-ALS, is well-suited to modeling multiple datasets

parametric model for elution profile: exponentially modified Gaussian (EMG)



[Global analysis]

idea of global analysis:

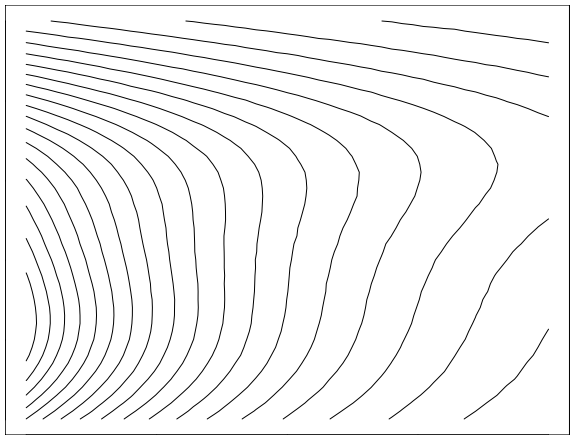
- formulate parametric model for elution profiles $C(\theta)$
- set starting values θ_0
- θ_0 allows solving for mass spectra as $S^T = C^+ D$
- this allows determination of residuals as $D - C(\theta_0)S^T = D - C(\theta_0)C^+ D$
- residuals are iteratively minimized with respect to θ
- this is instance of separable nonlinear least squares

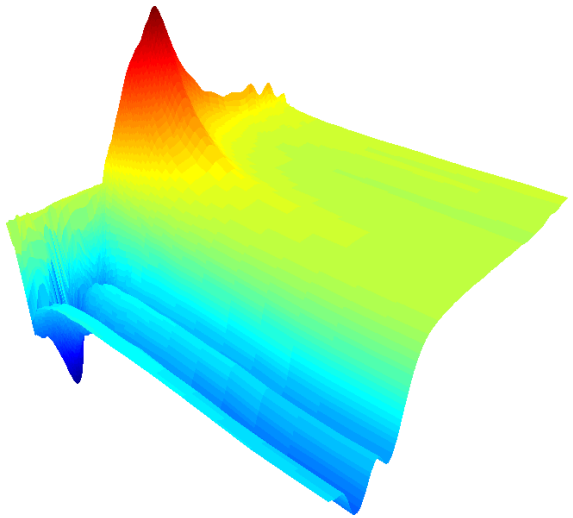
model for D is completely determined via relatively small number of parameters

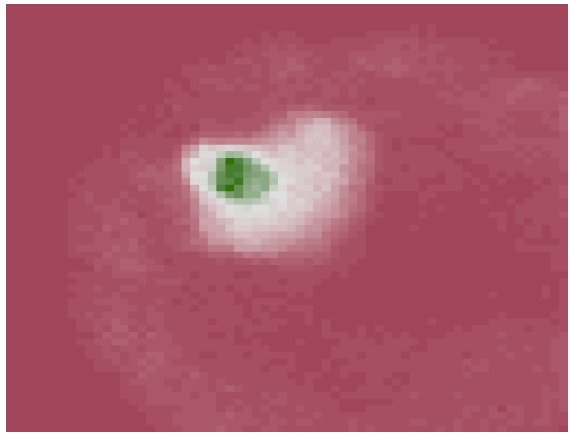
[Global analysis]

TIMP is a package for fitting separable nonlinear models that has been applied to measurements arising in

- time (and/or temperature, polarization, pH)-resolved spectroscopy
- fluorescence lifetime imaging microscopy (FLIM)







New options for fitting mass spectroscopy data have been recently added to **TIMP**: address outliers, baseline correction/estimation, saturation

[Comparison of methods]

	peak-based	MCR-ALS	global analysis
fast	yes	yes	no
integrated with mass spec. libraries	yes	no	no
model multiple datasets	no	yes	yes
resolve completely overlapping components	no	yes	yes

cons peak-based:

- broad elution profiles may be missed, or their local maxima may be detected as separate components

cons MCR-ALS:

- flexibility in shape of columns of C may allow more than one solution with same residual sum of squares

cons global analysis:

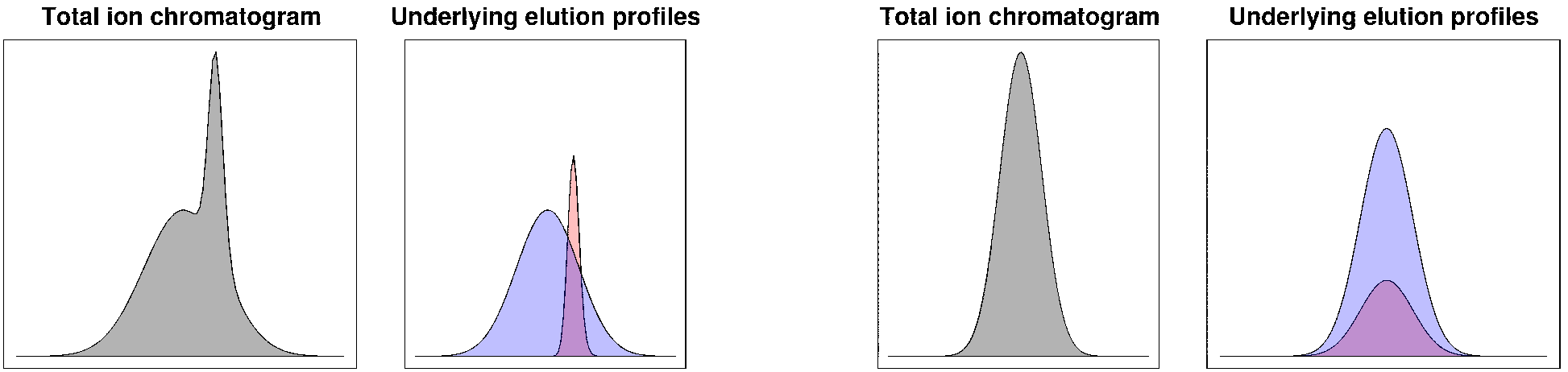
- requires user-intervention in model-specification

[Resolution of completely co-eluting components]

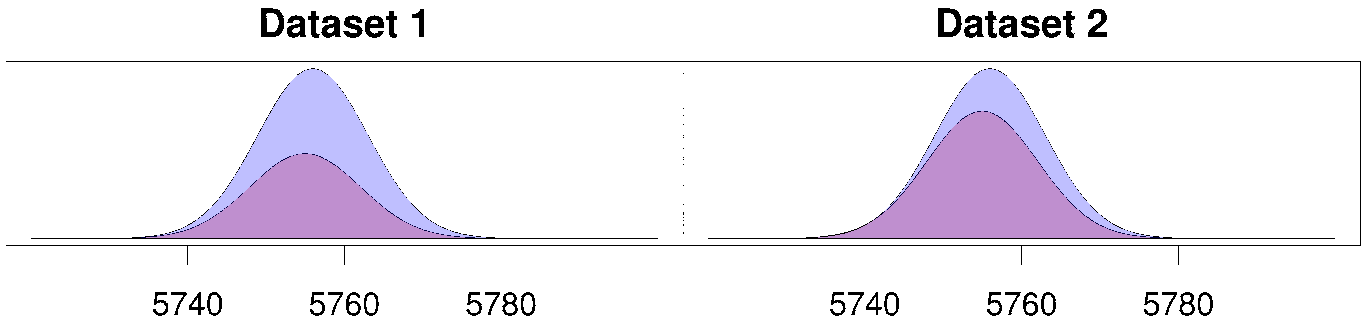
plots in color depict contributing elution profiles

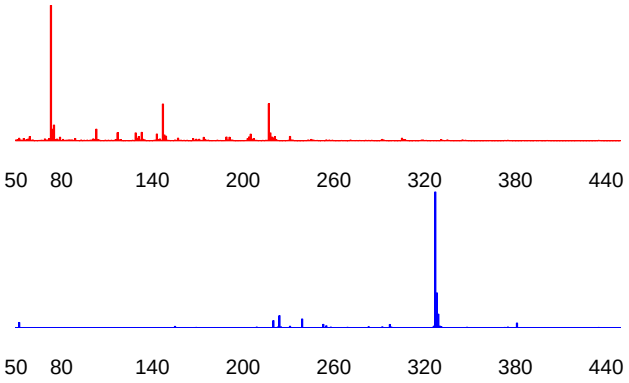
peak matching / MCR-ALS / global analysis can be used for component resolution when elution profiles are not completely overlapping (top row)

either MCR-ALS or global analysis make possible resolution of completely overlapping components (bottom row)



simulate data with Poisson noise, using elution profiles and mass spectra shown below

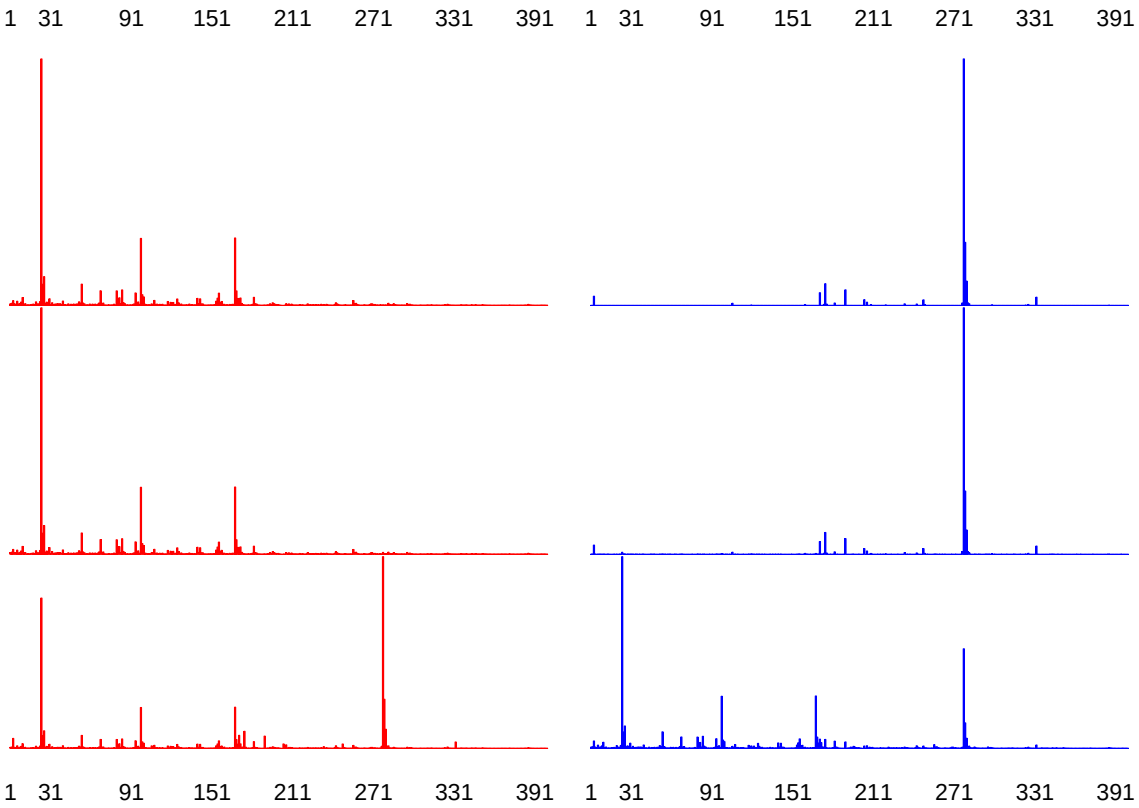




EMG parameters for elution profiles:

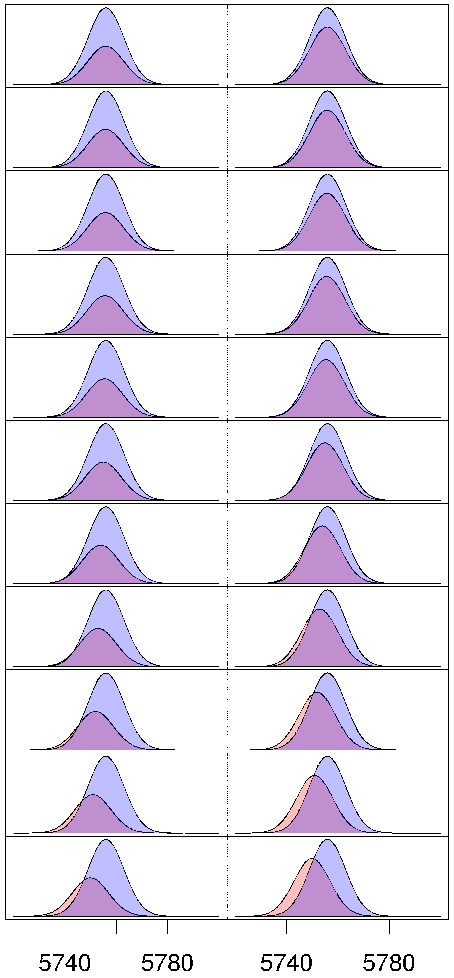
	component 1	component 2
location	5754	5755
FWHM	7	7
rate	1	1
amplitude, dataset 1	1	2
amplitude, dataset 2	1.5	2

[Resolution of completely co-eluting components]

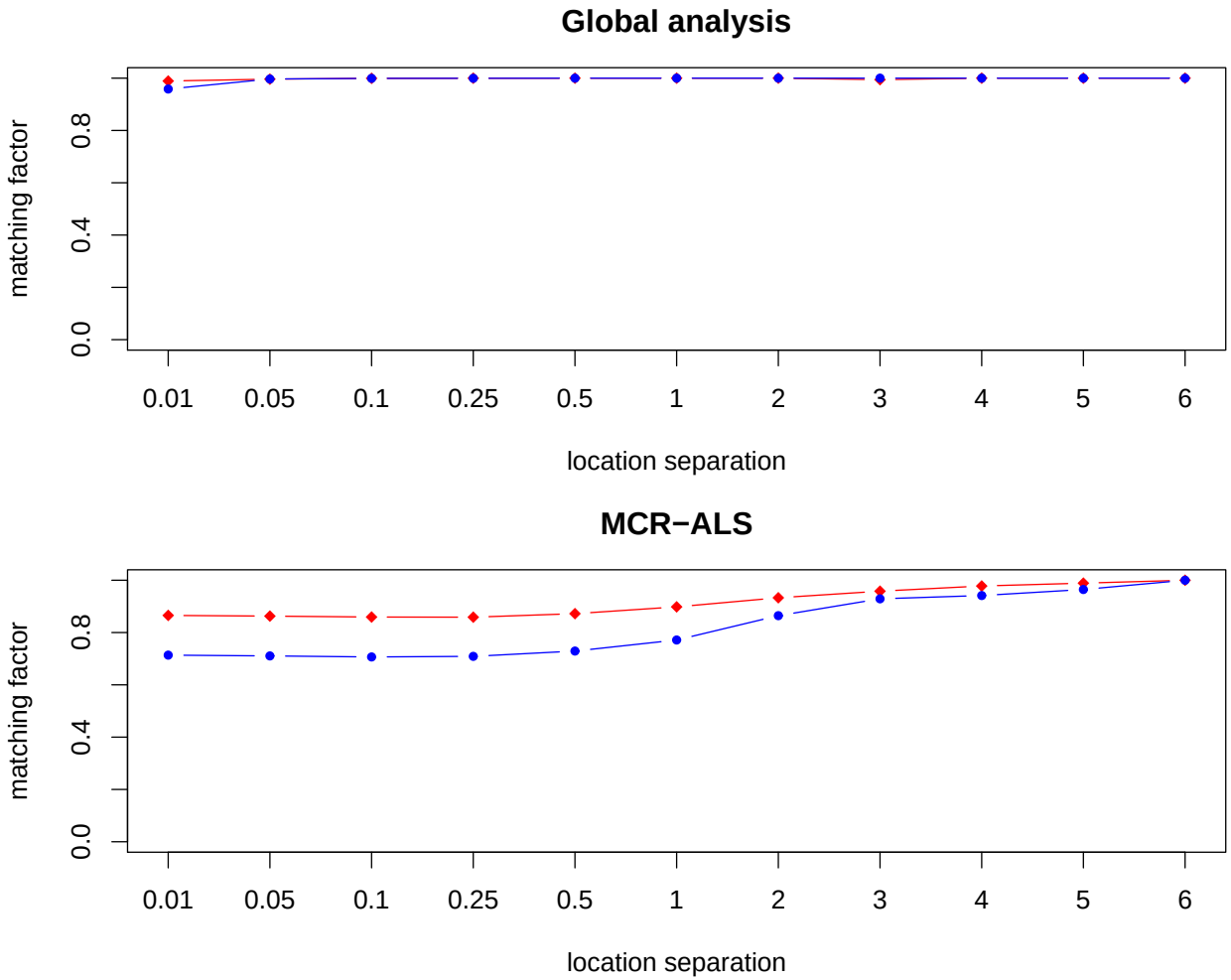


mass spectra estimated by global analysis (middle row)
match the spectra used in simulating the data (top row)
MCR-ALS (bottom row) returns a linear combination of the true mass spectra

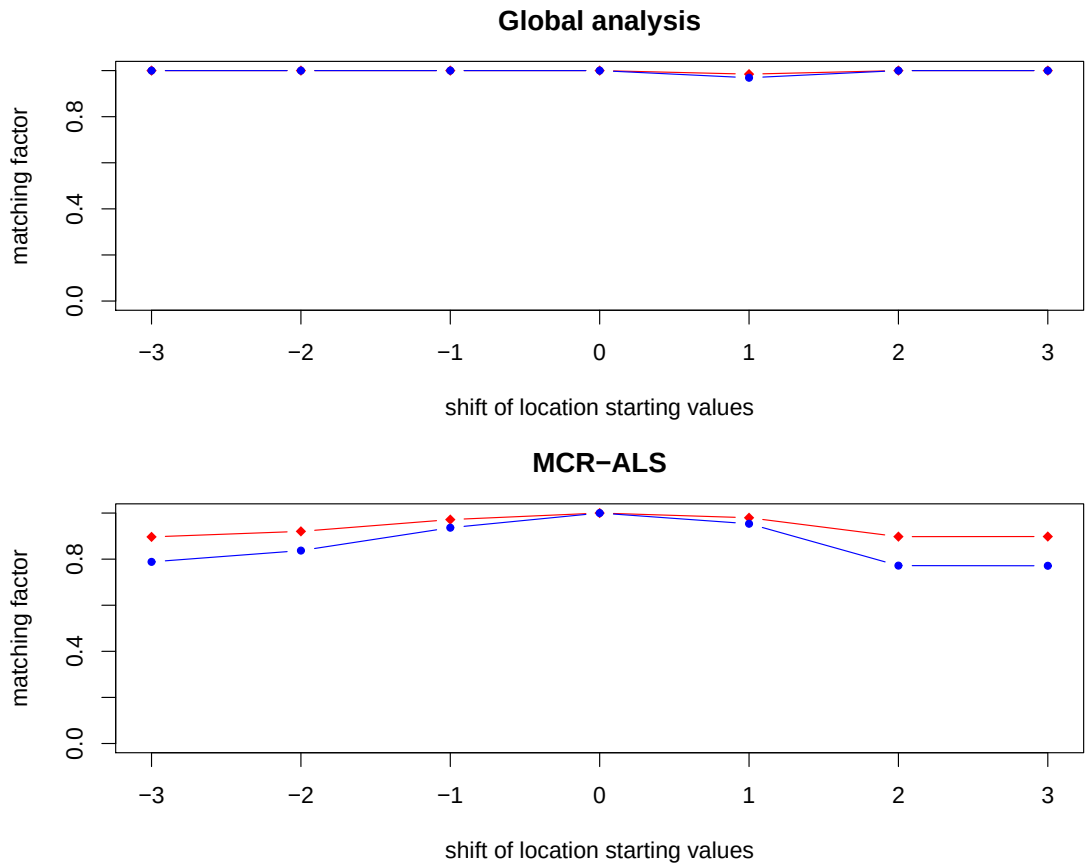
can examine many other problems with varying overlap of the elution profiles, from starting values for C that are slightly shifted versions of the values used in simulating the data:



observe that as location of elution profiles becomes more similar, MCR-ALS fails:



while MCR-ALS is not able to estimate good mass spectra from slightly wrong starting values, given perfect starting values it also performs well



A range of solutions have the same residual sum of squares under MCR-ALS; this ambiguity is a major drawback of the flexibility inherent to the method

Conclusions:

- component resolution problems in GC/MS and LC/MS data possible to address via
  - parametric global analysis: R package **TIMP**
  - non-parametric MCR-ALS: R package **ALS**
- both methods can deal with completely co-eluting components
- in certain situations global analysis returns better estimates of the mass spectra than MCR-ALS

Future work:

- a significant disadvantage of global analysis is run-time and user-intervention needed for model specification; for high-throughput deployment, more automation is necessary
- develop criteria to identify regions of data that would benefit from global analysis

[**Acknowledgments:**]

Ivo van Stokkum (Vrije Universiteit Amsterdam, Biophysics)

Linda Switzar (Vrije Universiteit Amsterdam, Analytical chemistry)

Filipe Lopes (Vrije Universiteit Amsterdam, Analytical chemistry)