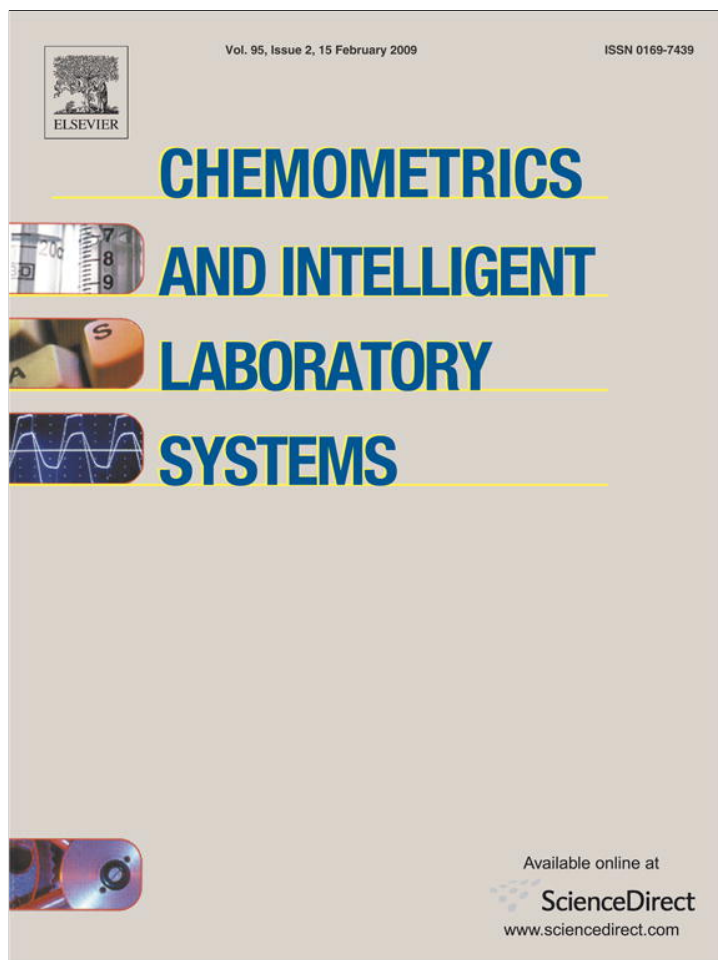




This article appeared in a journal published by Elsevier. The attached copy is furnished to the author for internal non-commercial research and education use, including for instruction at the authors institution and sharing with colleagues.

Other uses, including reproduction and distribution, or selling or licensing copies, or posting to personal, institutional or third party websites are prohibited.

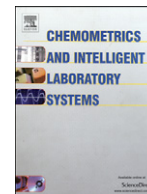
In most cases authors are permitted to post their version of the article (e.g. in Word or Tex form) to their personal website or institutional repository. Authors requiring further information regarding Elsevier's archiving and manuscript policies are encouraged to visit:

<http://www.elsevier.com/copyright>



Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Global analysis of multiple gas chromatography–mass spectrometry (GC/MS) data sets: A method for resolution of co-eluting components with comparison to MCR-ALS

Ivo H.M. van Stokkum^{a,*}, Katharine M. Mullen^a, Velitchka V. Mihaleva^b^a Department of Physics and Astronomy, Faculty of Sciences, Vrije Universiteit Amsterdam, de Boelelaan 1081, 1081 HV Amsterdam, The Netherlands^b Plant Research International, PO Box 16, Wageningen, 6700 AA, The Netherlands

ARTICLE INFO

Article history:

Received 24 July 2008

Received in revised form 29 September 2008

Accepted 1 October 2008

Available online 30 October 2008

Keywords:

Multivariate curve resolution alternating least squares

Global analysis

Mass spectrometry

Co-elution

ABSTRACT

Global analysis has been applied to resolve components in multiple gas chromatography–mass spectrometry (GC/MS) data sets. Global analysis methodology is based upon a parametrized model of the observed data, including random (and possibly also systematic) errors. Each elution profile is described as a function of a small number of parameters. We successfully based the description of elution profiles on an exponentially modified Gaussian. The mass spectra were described non-parametrically. Model usefulness is judged by the quality of the fit and whether the estimated parameters that describe the elution profiles and mass spectra of components are physically interpretable. Advantages of the method are most evident with multiple data sets and overlapping elution profiles. Differences between data sets are described by alignment parameters and by relative amplitude parameters. The estimated mass spectrum is identical between experiments. Global analysis and multivariate curve resolution alternating least squares (MCR-ALS) are the only methods currently developed for component resolution for the case of completely co-eluting compounds in mass spectrometry data. In the present contribution global analysis is shown to have better performance than MCR-ALS in terms of the estimated mass spectra for a variety of simulated GC mass spectrometry datasets representing components that are completely co-eluting.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

The component resolution problem for an $m \times n$ matrix of data D_1 can be stated as the problem of estimating the matrices C_1 and S from D_1 in

$$D_1 = C_1 S^T \quad (1)$$

such that the matrix C_1 is $m \times n_{\text{comp}}$, and each column represents the evolution of a component of D_1 in the variable with which the rows of D_1 are resolved, and such that the matrix S is $n \times n_{\text{comp}}$, and each column represents the component of D_1 in the variable with which the columns of D_1 are resolved. For chemistry applications the components often correspond to the different chemical compounds in the sample underlying the measurements, and the problem is sometimes referred to as *deconvolution*. Usually it is assumed that the noise applied to D_1 is comprised of independent identically distributed Gaussian processes and C_1 and S are estimated under least squares criteria.

It is often of interest to perform component resolution of many related data matrices $D_1, \dots, D_K$ simultaneously, where the datasets are related in that they represent components with the same properties in

the variable with which the columns of D_i are resolved, giving rise to the equation

$$\begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_K \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_K \end{bmatrix} S^T \quad (2)$$

By analogues of the Beer–Lambert law, the linear relation contained in Eqs. (1) and (2) well-describes many varieties of two-way data arising in spectroscopy and spectrometry experiments. The stochastic element of the model is neglected in Eqs. (1) and (2), but is often assumed to be comprised of additive Gaussian white noise, as we will assume throughout.

Algorithms to solve the component resolution problems as formulated above may be roughly categorized as those that use constraints but no parametric model, and those that employ a parametric (that is, functional) description for some aspect of the data. The former category of algorithms is often termed self-modeling curve resolution (SMCR) techniques, and has been recently reviewed [1]. An important self-modeling curve resolution technique that we will concentrate on here is multivariate curve resolution alternating least squares (MCR-ALS) [2–4], which has a long and successful history of application to a variety of analytical problems [5].

The class of algorithms that employ a parametric model for some aspect of the data includes global analysis algorithms that describe

* Corresponding author.

E-mail address: ivo@nat.vu.nl (I.H.M. van Stokkum).

each component of C_i in terms of a nonlinear function, and treat the entries of S as conditionally linear parameters that are not described functionally. Global analysis methods have been widely applied to component resolution problems in spectroscopy [6–8] and microscopy [9–11] applications, where the optimized parameters of the functional description of C_i are interpreted physically, allowing insight into the dynamical processes underlying the data. Methods that are hybrid in the sense that they use MCR-ALS to iteratively improve estimates for $C_1, \dots, C_K$ and S , but refine the MCR-ALS estimates for $C_1, \dots, C_K$ by fitting them with a parametric model each time $C_1, \dots, C_K$ is updated have also been investigated [12,13].

In this study, we will consider MCR-ALS and global analysis for component resolution problems in mass spectrometry data. Whereas MCR-ALS has a well-developed history of successful application in this domain [14–17], software tools for the application of global analysis methods to mass spectrometry data are described here for the first time. Section 3 shows via several case studies a proof-of-concept of the utility of global analysis for component resolution problems arising in gas chromatography–mass spectrometry data.

Mass spectrometry data associated with a particular sample is often stored as a matrix D_i resolved with respect to time and mass-to-charge ratio (m/z), so that each data point is a count of the number of molecular fragments having a given m/z measured at a given time. Then C_i represents elution profiles of components in time and S represents mass spectra. The aim of component resolution in the context of mass spectrometry datasets $D_1, \dots, D_K$ is to determine the mass spectra of components and, possibly, obtain the retention times of components, where the retention time is defined as the time at which an elution profile is maximal. Estimated mass spectra and retention times are matched against mass spectra and retention times of known compounds stored in a library [18], allowing the compounds represented by the estimated spectra to be identified. The relative concentration of a particular component k in the samples represented by datasets D_i and D_j may be estimated by the quotient between the area of the estimated elution profiles of the component, that is, by $C_i[k, k]/C_j[k, k]$.

In the case that the elution profiles of two or more components are overlapping, the components are said to be *co-eluting*, as in Fig. 1. If the profiles overlap exactly and have the same shape, as in the bottom panel of Fig. 1, only by analyzing several datasets simultaneously in

which the amplitude of the profiles varies are the components possible to resolve. Much attention has been paid to the resolution problem in the case of co-eluting components [3,17]. MCR-ALS is currently considered to be the unique method capable of resolving components in the case of completely overlapping elution profiles [14]. We present global analysis as an alternative component resolution methodology for the case that compounds are completely co-eluting, and make available open source implementations of MCR-ALS and global analysis in the form of packages for the R language and environment [19], so that others may reproduce the examples discussed here and possibly extend the methodology.

The remainder of this study is organized as follows: Section 2 introduces the global analysis methodology for the analysis of time-resolved spectroscopy data. Section 3 shows several applications of global analysis to component resolution problems in measured gas chromatography–mass spectrometry (GC/MS) data. Section 4 gives an overview of the non-parametric MCR-ALS methodology for component resolution problems. Section 5 compares the performance of MCR-ALS and global analysis in terms of the matching factor of estimated mass spectra in a simulation study inspired by the GC/MS data discussed in Section 3. Sections 6 and 7 contain discussion and conclusions, respectively.

2. Parametric model-based global analysis

The recovery of $C_1, \dots, C_K$ and S from $D_1, \dots, D_K$ can be approached via global analysis methods based on a parametric model [8]. In the mass spectrometry context, a parametric model is applied to the description of the elution profiles $C_1, \dots, C_K$, so that each elution profile in each dataset is described by a peak function such as the exponentially modified Gaussian (EMG). The process of model fitting optimizes the parameters describing the elution profiles and treats the mass spectra as conditionally linear on their estimates. Usually the criterion optimized is the sum of squared differences between the model evaluated at the parameter estimates and the data. In this case the model fitting problem is an instance of separable non-linear least squares, which is possible to address with the variable projection algorithm [20]. In mass spectrometry applications the data represent counts and can be expected to behave as a Poisson distributed stochastic process. Since the number of counts is usually large, minimization of the actual log-likelihood function would offer very little improvement over optimization of the sum of squared deviations [21].

2.1. Functions for the description of elution profiles

Many functions are possible to apply to the description of chromatographic peaks [22]. The exponentially modified Gaussian (EMG) function is the most popular in practice and can be used to describe a variety of peak shapes with a relatively small number of parameters. It employs three parameters, location λ , full width at half-maximum (FWHM) Δ , and rate k to determine each elution profile c over a vector of times t as

$$c(t, \lambda, \Delta, k) = \frac{\exp(-kt)}{2} \exp\left(k\left(\lambda + k \frac{\Delta^2}{2}\right)\right) \left\{1 + \operatorname{erf}\left[\frac{t - (\lambda + k \frac{\Delta^2}{2})}{\sqrt{2}\Delta}\right]\right\} \quad (3)$$

where $\tilde{\Delta} = \Delta / (2\sqrt{2 \log(2)})$. With a positive decay rate k the elution profile exhibits a tail. To describe fronting, a negative decay rate k can be used, and the time argument is reversed.

Thus the model for D_1 representing n_{comp} components reads

$$D_1(t, \mu) = \sum_{l=1}^{n_{\text{comp}}} c(t, \lambda_l, \Delta_l, k_l) S_l(\mu) \quad (4)$$

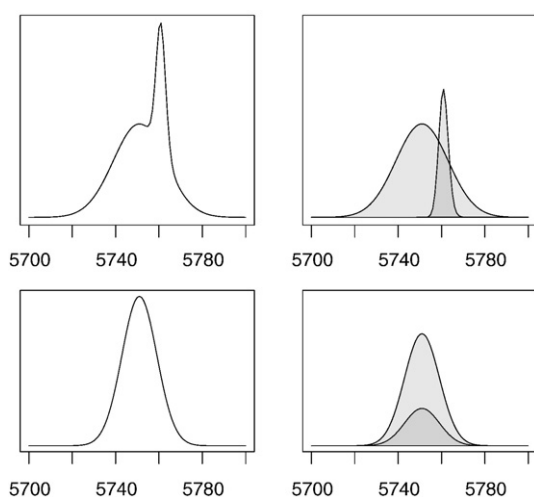


Fig. 1. The plots on the left depict slices of a chromatogram in which overlapping elution profiles are represented, shown in the plots on the right. The elution profiles in the top plots are overlapping, but not completely so, allowing a variety of algorithms to be applied for component resolution. The plots on the bottom involve completely overlapping components, and only by simultaneously analyzing datasets in which the amplitudes of the components vary is resolution possible, with either MCR-ALS or global analysis.

where μ means the m/z axis. For an additional experiment p we assume that the shape parameters k_l and Δ_l are independent of experiment, and introduce alignment parameters λ_p (assumed to be identical for all components), and amplitude parameters $\alpha_{l,p}$, yielding a model parametrized as

$$D_p(t, \mu) = \sum_{l=1}^{n_{\text{comp}}} c(t, \lambda_l + \lambda_p, \Delta_l, k_l) a_{l,p} S_l(\mu) \quad (5)$$

When the shape varies with experiment, or when the alignment varies between components because of interaction, additional parameters can easily be introduced.

It is usually desirable to keep the total number of parameters describing the elution profiles as low as possible, which renders the model better determined and faster to optimize. To address the problem of intensity ambiguity, one of the amplitude parameters is fixed for each component. The amplitude of each component is fixed in one dataset, so that the other amplitudes describing the component are relative to the fixed value and the model is well-determined. Provided all components are present in the D_1 , it is often convenient to fix the amplitude parameters of all components in this dataset to 1, while allowing all amplitude parameters in datasets $D_2, \dots, D_K$ to remain free.

2.2. Data selection and pre-processing

As in MCR-ALS [14], for global analysis the data is cut into time windows so that $D_1, \dots, D_K$ represent on the order of 1–10 components, as described in Section 4.2. Instead of correcting for a baseline in the data, which can be problematic in the case that its contribution is changing in time, global analysis allows the possibility of modeling the baseline term as a component, assigned its own mass spectrum. It is also possible to allow the parameter describing the location of each elution profile to vary per-dataset, meaning that pre-alignment is not a prerequisite for data analysis.

2.3. Number of components and starting estimates

The estimation of the number of components proceeds along the same lines as for MCR-ALS, as described in Section 4.3.

Starting estimates for global analysis may be estimated by obtaining starting estimates as for MCR-ALS as described in Section 4.4, and then fitting the desired functional description of the elution profiles to these estimates. The decision regarding which parameters to make common between elution profiles in different datasets is at present made by hand, though automation would be desirable. For numerical reasons, the starting values should not result in peaks associated with precisely the same parameter values (shifting the location of one peak slightly resolves any problems).

2.4. Parameter estimation

For each component $c_l(t)$ is described by an EMG function with three parameters k_l , λ_l and Δ_l . Each additional experiment p requires one time shift parameter λ_p for alignment, and n_{comp} amplitude parameters $a = [a_{1,p}, a_{2,p}, \dots, a_{n_{\text{comp}},p}]$ representing the amplitude of each component, so that we end up with $(2 + n_{\text{exp}})n_{\text{comp}} + n_{\text{exp}} - 1$ intrinsically nonlinear parameters for n_{exp} experiments. The mass spectra S contain $n_{\text{comp}}n_\mu$ parameters, where n_μ is the number of masses represented by the data, so that the number of parameters represented by S is typically on the order of 10^3 . These parameters, however, are conditionally linear, and can be eliminated analytically from the problem using the variable projection method reviewed by [20]. Note that in this way the model fitting process proceeds much more efficiently. Since negative values of S cannot be interpreted, these parameters are constrained to nonnegative values. The in-

corporation of nonnegativity constraints on the conditionally linear parameters S within global analysis is described by [23,24] and is implemented using a nonnegative least squares (NNLS) algorithm by Lawson and Hanson [25,26].

A background contribution may be assumed to be constant in the time window analyzed (and has only a mass spectrum, but no parameters). More complex formulations of the background contribution are possible by introduction of a slope parameter or some other more flexible parametric description.

2.5. Residual analysis

The application of global analysis to measured data described in Section 3 uses the following methodology to analyze the residuals, detect outliers, and deal with saturation effects.

The residual matrix is analyzed using a singular value decomposition (SVD):

$$D_{\text{res}}(t, \mu) = D(t, \mu) - D_{\text{fit}}(t, \mu) = \sum_{i=1}^m u_i(t) s_i w_i^T(\mu) \quad (6)$$

where u_i and w_i are the left and right singular vectors, s_i are the sorted singular values, and m is the minimum of the number of rows and columns of the data matrix. When multiple experiments are simultaneously analyzed the residual matrices are concatenated. Extra measures are needed to identify residuals that are caused by detector saturation, and to identify outliers.

A residual caused by detector saturation is defined by $D_{\text{res}}(t, \mu) > \alpha_1 D_{\text{fit}}(t, \mu)$ and $D(t, \mu) > \alpha_2$, where α_1 and α_2 are values that depend upon the experiment. This definition depends upon D_{fit} and thus must be determined iteratively.

To identify outliers all residuals at a particular m/z value of experiment p are sorted, and the Lower and Upper Fourth are determined, and the Fourth Spread. Then the Lower Limit (LL) and Upper Limit (UL) are defined as

$$\text{LL} = \text{Lower Fourth} - \alpha \cdot \text{Fourth Spread} \quad (7)$$

$$\text{UL} = \text{Upper Fourth} + \alpha \cdot \text{Fourth Spread} \quad (8)$$

The data points associated with residuals outside these Lower and Upper Limits are identified as outliers [27]. Experimentally we found that a typical value of α useful in the context of GC/MS data is 5. A low weight is applied to huge signals that can saturate the detector e.g. the trimethylsilyl (TMS) peak at m/z 73. Likewise, low weights are applied to outliers and residuals caused by detector saturation (determined by the above procedures). High weights can be applied to characteristic masses provided the signal-to-noise ratio is sufficient. Instrument nonlinearity is currently neglected, but in principle it could be included in the model function. The global analysis method developed here was applied to two large sets of data, namely tomato data sets described in [28] consisting of 37 different tomato fruit varieties, and standard mixture data described in [14] comprising samples with up to 101 different chemical compounds. No preprocessing was applied to the data, except for a baseline correction when needed.

2.6. Refinement

After a global analysis a further refinement is possible. The estimated mass spectra can be used in a weighted NNLS fit for each time point of each dataset p to estimate the amplitudes of the concentrations, so that c_t in $D_{\text{pt}}^W = S^W c_t$ has solution

$$c_t = \text{NNLS}(S^W, D_{\text{pt}}^W) \quad (9)$$

where the superscript W is used to indicate that weights are applied, and D_{pt}^W is a column vector of all masses at time t of experiment p . S^W

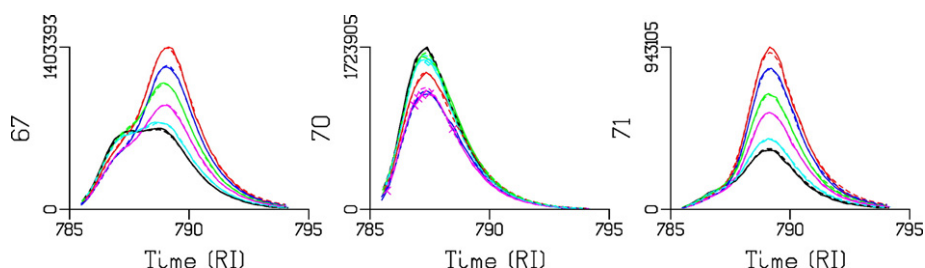


Fig. 2. Tomato I. Representative traces (aligned with estimated shift parameters) below saturation threshold at m/z values of 67, 70 and 71. Colors indicate the six different data sets, solid lines are data and dashed lines indicate fit. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

is the weighted matrix of estimated mass spectra. The refinement provides an opportunity to check the assumed peak shape and peak resolution.

2.7. Implementation

The global analysis method discussed here is publicly available under the terms of the GNU General Public License in the form of the R package **TIMP** [29], from the Comprehensive R Archive Network (CRAN) [19].

2.8. Identification

The mass spectra estimated via global analysis in Section 3 were searched against the NIST05 library [30] extended with the Golm mass spectral library [31] of derivatized compounds. For each spectrum 10 hits were retrieved. The matching factors reported by the MS Search 2.0 program [30] are between 0 (no match) and 1000 (perfect match). Two different values for the matching factors, MF and RMF, are reported. In the calculation of MF the experimental spectrum is used as a template, whereas for RMF the template is the library spectrum. Comparable MF and RMF values indicate that there are no additional and/or missing values in either of the spectra. To increase the reliability of the identification we have included the retention index (RI)

in the evaluation of the library hits. However, RI data are available for only a small fraction of the compounds in the NIST05 library. Therefore we have used a recently developed quantitative structure-retention index model [32] for the estimation of RI based upon the structure of the compounds.

3. Application of global analysis to measured GC/MS data

We successfully performed over 100 case studies on selected time windows using different amounts of datasets, from which we present here four illustrative examples. The first and second case study presented in Section 3.1 are referred to as Tomato I and Tomato II, respectively, while the first and second case study described in Section 3.2 are referred to as Mixture I and Mixture II, respectively.

3.1. Tomato data

The case studies here demonstrate the ability of global analysis to resolve two overlapping components and resolve correct mass spectra in the case of saturation effects (case study Tomato I) and the ability to resolve six overlapping peaks (case study Tomato II).

In case study Tomato I, we use six tomato samples (e.g., datasets) and focus on a small part of the chromatogram where two components clearly overlap. At m/z values of 70 and 71 the peak location is below RI 787 and above RI 788, respectively, indicating that these are characteristic masses, whereas at an m/z value of 67 both components contribute almost equally. These six data sets (shown at selected m/z values in Fig. 2) were simultaneously analyzed, resulting in the mass spectra of Fig. 3 (Top). Both mass spectra were successfully identified as elaborated in Table 1. The mass spectrum at RI 786.6 was identified as 3(Z)-Hexenal. The first five hits retrieved from the library had very similar MF and RMF values but their RI's differed. Using both the matching factors and RI, 3(Z)-Hexenal was selected as the best hit. The component at RI 788.5 was identified as Hexanal. There was a much larger difference between the MF values of Hexanal and the rest of the hits. Also, the predicted RI was close to that estimated in the experiment. These two compounds have been previously identified in tomato samples [28,33].

The estimated FWHM parameters are 1.6 and 1.7 RI units, and the locations of these two peaks are somewhat more than one FWHM

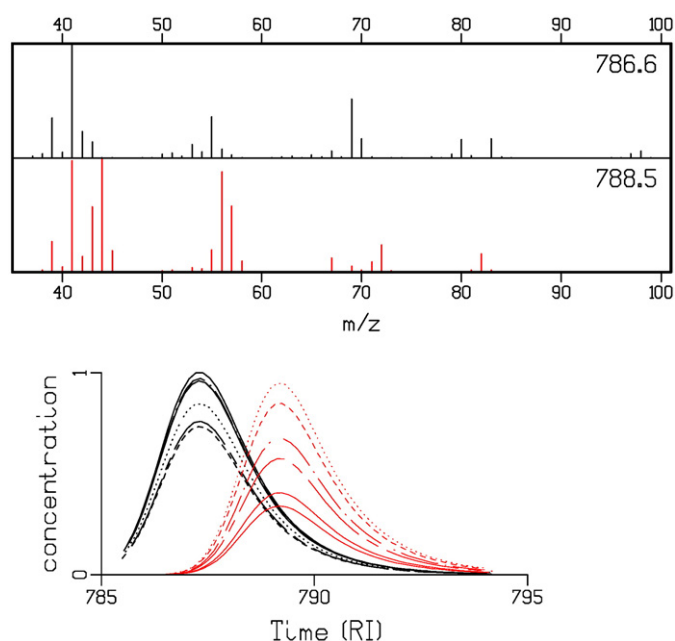


Fig. 3. Tomato I. (Top) Normalized mass spectra with location λ of peak maximum indicated in the upper right corner. (Bottom) Estimated elution profiles with linetypes indicating the different experiments. The contribution of each component to each data set is the product of each elution profile times the associated normalized mass spectrum times the scaling factor 10,259,207.

Table 1
Identifications for Tomato I and II case studies described in Section 3.1

Name	RI	MF	RMF	RI _{lit}	RI _{pred}
3(Z)-Hexenal	786.6	842	843	799	813
Hexanal	788.5	901	902	784	827
Benzyl Alcohol	1039.4	898	902	1026	1006
3,5-Octadien-2-ol	1040.6	723	738	–	1095
5-Methyl-3-hepten-2-one	1040.6	716	758	–	966
Chloro-2-hydroxy-propanoic acid	1041.1	613	643	–	1081
5-Ethyl-2(5H)-Furanone	1042.0	844	865	952	1004
2-Isobutylthiazole	1042.8	841	861	1013	1122
1-ethyl-4-methyl-Benzene	1043.3	598	711	952	1009

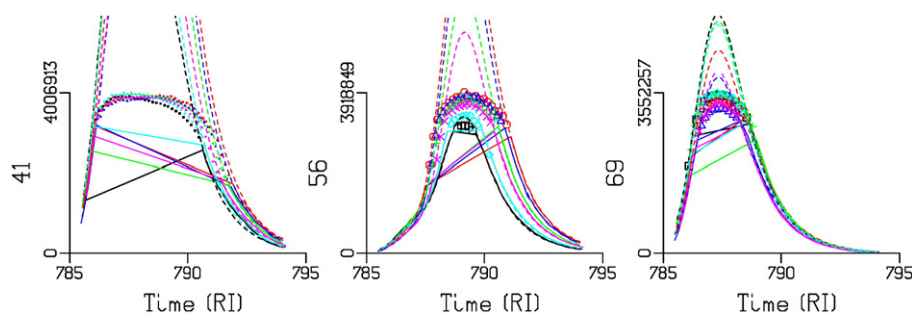


Fig. 4. Tomato I. Representative traces (aligned with estimated shift parameters) above saturation threshold at m/z values of 41, 56 and 69. Colors indicate the six different data sets, solid lines are data and dashed lines indicate fit. Symbols indicate saturation data points. The two tails of the peak are connected by straight lines. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

apart. The estimated decay rate parameters are both about 0.8/RI, indicating appreciable tailing. These components could not be reliably resolved by analysis of individual data sets using e.g. the Automated Mass Spectral Deconvolution and Identification System (AMDIS) program [34]. The estimated elution profiles are depicted in Fig. 3 (Bottom). Note that the amount of component present in the data sets varies appreciably, as indicated by the different linetypes.

Examples of saturation are depicted in Fig. 4. Note that at these masses most of the peak is clipped, and it is essential to reconstruct the original shape using the unclipped tails of the elution profile. In this way a reliable estimate of the mass spectrum can also be made at these critical masses. It turns out that this is essential for adequate identification of the compounds present in the samples. From a further in depth analysis of the residuals (visible as systematic misfit at e.g. the μ value of 41) we infer that a small instrument nonlinearity may be present. In this way, these two overlapping components have been successfully analyzed in 37 different tomato varieties simultaneously.

In the case study Tomato II we present an in-depth analysis of a much smaller peak around RI 1042 in six tomato samples.

Note the large differences in the shapes of the elution profiles in Fig. 5, and also the differences in the amplitudes of the datasets, as indicated by the different colors. E.g., at m/z values of 58 and 99 the cyan and red are largest, peaking near RI 1043, whereas at 57 green and blue dominate, peaking somewhat earlier, and at 105 blue and black dominate. To describe these differences six components were needed to fit the data, whose mass spectra are shown in Fig. 6 (Top). In this region Tikunov et al. [28] have identified the components at RI 1039.4 and 1042.8 as Benzyl alcohol and 2-Isobutylthiazole. These compounds were found as the best hits also in our analysis (see Table 1). The component at RI 1042.0 was identified as 5-Methyl-2

(5H)-furanone, a compound recently found to be present in small amounts in tomato [35]. The best hit for the component at RI 1041.2 was that of a chlorinated acid. Both the MF and RMF values were very low which makes the identification doubtful. Also, only a few chlorinated compounds have been identified in tomato [33]. Unsaturated alcohol (3,5-Octadien-2-ol) and ketone (5-Methyl-3-Hepten-2-one) had very similar MF and RMF values with the mass spectra at RI 1040.6. The RI's were also close to the experimental value. These two compounds have not been identified in tomato but other unsaturated alcohol and ketones are known to be present. There was a large difference between the MF and RMF values for the hits of the mass spectrum at RI 1043.3. This is an indication that the extracted mass spectrum is not pure probably due to the small contribution of the component. The hit list for this component contained compounds with very different RI's. The best hit based on RI was for 1-Ethyl-4-methyl-benzene, a compound known to be present in tomato [33].

The estimated elution profiles of the components are depicted in Fig. 6 (Bottom). The estimated FWHM parameters are again around 1.6 RI units, thus the locations of these six components are within a range of 2.5 FWHM.

Note that the contributions of the fifth and sixth component (indicated by cyan, and black) are very small, as evident from Fig. 6 (Bottom). Nevertheless they can be reliably estimated from these six tomato varieties.

3.2. Standard mixture data

Two benefits of global analysis will be illustrated in turn with case studies Mixture I and Mixture II on standard mixture data from [14]. The data were previously analyzed in [14] with a self-modeling

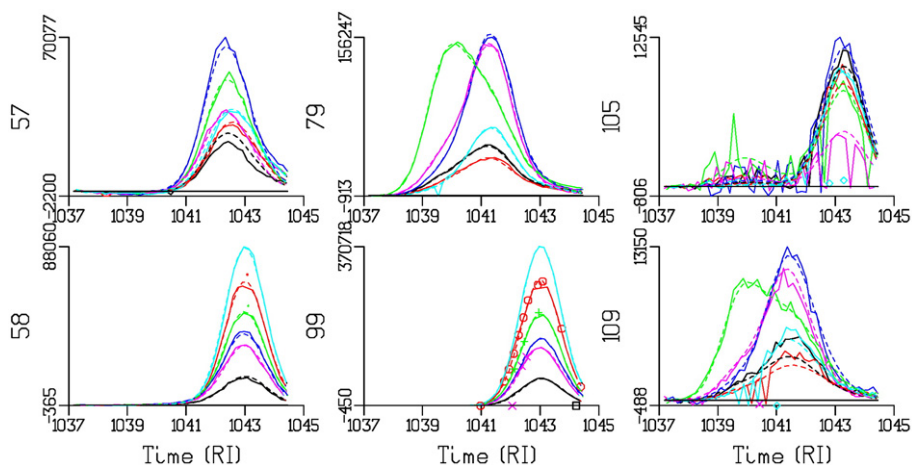


Fig. 5. Tomato II. Representative traces (aligned with estimated shift parameters) at m/z values between 57 and 109. Colors indicate the different data sets, solid lines are data and dashed lines indicate fit. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

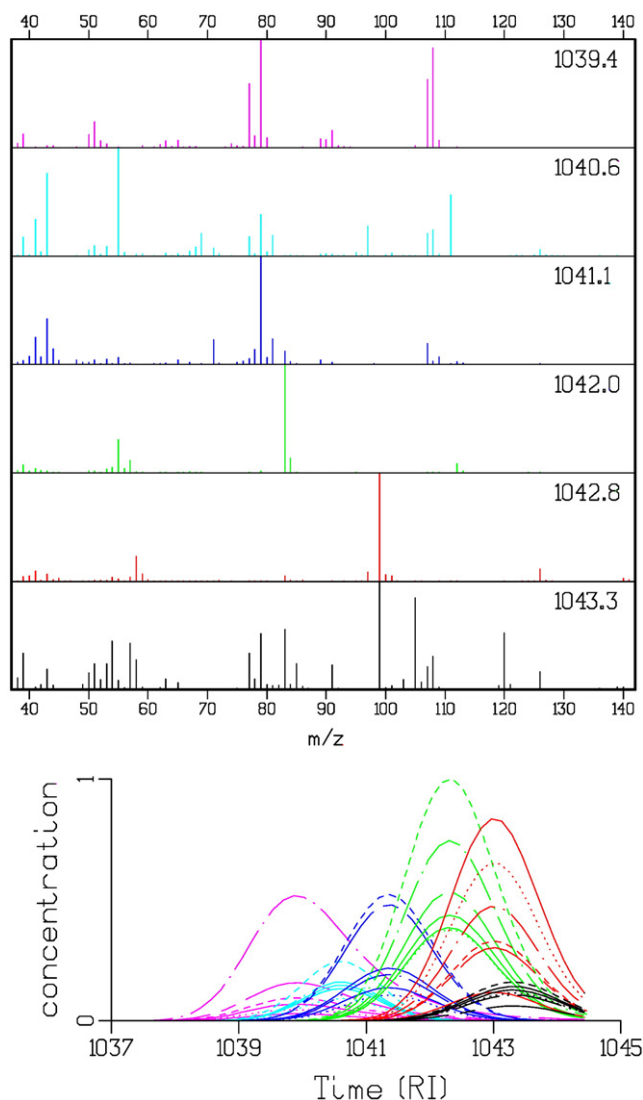


Fig. 6. Tomato II. (Top) Normalized mass spectra with location λ of peak maximum indicated in the upper right corner. (Bottom) Estimated elution profiles with colors indicating the six components, and linetypes indicating the different experiments. The contribution of each component to each data set is the product of each elution profile times the associated normalized mass spectrum times the scaling factor of 519,121. In order to improve visibility of the small concentrations, we have used a square root scaling of the maxima. Thus e.g. the black concentration maximum at about 0.1 corresponds to a true maximum of about 0.01. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

method, hierarchical MCR-ALS. The example in Mixture I shows the ability of global analysis to resolve components with small amplitude relative to a large peak nearby. The ability of global analysis to resolve highly overlapping components is illustrated by the example in Mixture II. Both examples will proceed using only a small amount of data (6 or 3 data sets), whereas [14] used a large amount of designed data (62 data sets) that are analyzed simultaneously. The data are designed in that the concentration of some components is made to vary between datasets. Four representative traces from six standard mixture datasets are depicted in Fig. 7. Note that the peak at $m/z=327$ is about 200 times smaller than the maximum of the next peak at m/z values of 73 and 219. Furthermore the presence of the baseline aggravates this resolution problem.

The red mass spectrum in Fig. 8 (Top) was identified as trans-caffeic acid-3TMS (MF 949, RMF 949), and the green baseline spectrum shows TMS related peaks near 73 and 147. The blue and black mass spectra could not yet be identified. Although hits with relatively good matching factors were retrieved, the library RI's for

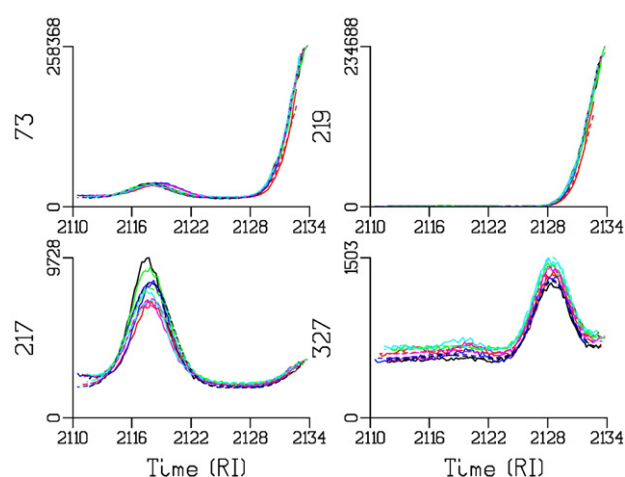


Fig. 7. Mixture I. Representative traces (aligned with estimated shift parameters) at m/z values between 73 and 327. Colors indicate the six different data sets, solid lines are data and dashed lines indicate fit. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

these compounds did not match the experimental RI. The global analysis algorithm with nonnegativity constraints on the mass spectra results in estimates for the black mass spectrum that are zero at masses below 217. Apparently it was not possible to estimate its small

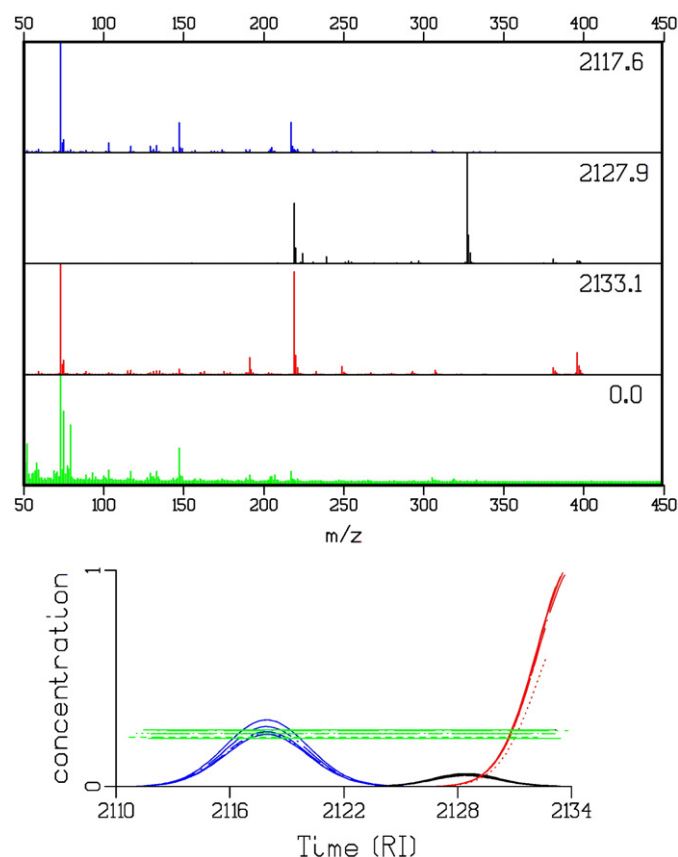


Fig. 8. Mixture I. (Top) Normalized mass spectra with location λ of peak maximum indicated in the upper right corner. Zero is the baseline spectrum. (Bottom) Estimated elution profiles with colors indicating the four components, and linetypes indicating the different experiments. The contribution of each component to each data set is the product of each elution profile times the associated normalized mass spectrum times the scaling factor of 243,828. In order to improve visibility of the small concentrations, we have used a square root scaling of the maxima. Thus e.g. the black concentration maximum at about 0.05 corresponds to a true maximum of about 0.0025. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

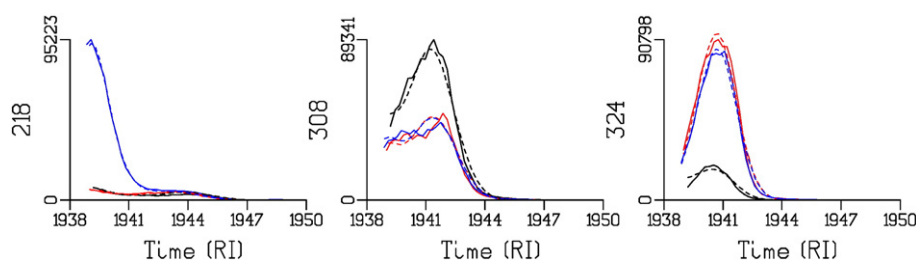


Fig. 9. Mixture II. Representative traces (aligned with estimated shift parameters) at m/z values of 218, 308 and 324. Colors indicate the three different data sets, solid lines are data and dashed lines indicate fit. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

amplitude (relative to the huge caffeic acid peak) near e.g. $m/z=73$. The estimated elution profiles of the components depicted in Fig. 8 (Bottom) highlight the small black component contribution.

The analysis for the Mixture I case study concerns the region from RI 2110–2134. Likewise we analyzed all bands in the region from Retention Index 1902–2433 in small time windows, using only 3 or 6 of the datasets simultaneously. Compared to hierarchical MCR-ALS methodology as described in [14], global analysis resolves many more small components using only a very limited amount of data. The Mixture II

case study concerns a particularly difficult overlap region from Retention Index 1938–1945. Five different components are present, with two pairs almost completely overlapping. In the three different data sets the amounts of these overlapping components were different by experimental design, thus allowing to resolve them (analogous to the natural component variability with tomato varieties).

Note the large differences in the shapes of the elution profiles at the three characteristic masses in Fig. 9, and also the differences in the amplitudes of the datasets, as indicated by the different colors. The estimated mass spectra shown in Fig. 10 (Top) were successfully identified, and the matching factors are shown in Table 2. Strikingly, just outside of this region two more mass spectra were identified as galacturonic acid at RI 1930 (MF 808/RMF 829) and RI 1953 (MF 862/RMF 875). [14] only reported two galacturonic acid peaks at RI 1940.2 and 1956.5. The difference between our RI 1953 and their RI 1956.5 can be well explained by absolute alignment uncertainty. They also report trans-p-coumaric acid at 1943.5 and coniferylalcohol at 1942.8, which is also about 3 RI units higher. Alignments of 3 RI units between their samples are common. We currently have no explanation for the two extra putative galacturonic acid peaks at RI 1930 and RI 1942.8. The estimated elution profiles of the components depicted in Fig. 10 (Bottom) demonstrate again the importance of amplitude variability for the resolution of overlapping components. Note also that the solid red elution profile is delayed relative to the dotted and dashed one.

3.3. Refinement of results

To further check the adequacy of the assumed EMG peak shape, we estimated concentration profiles according to the refinement described in Section 2.6. The profiles in Fig. 11 are very similar to the EMG shapes in Figs. 3 (Bottom), 6 (Bottom), 8 (Bottom), and 10 (Bottom). The improvement in the weighted root mean square error of the fit was 36% and 15% in case studies Tomato I and II, and 29% and 58% in case studies Mixture I and II. The green elution profiles of Mixture II deviate most from an EMG peak shape.

4. Multivariate curve resolution alternating least squares (MCR-ALS)

While MCR-ALS has been presented many times in the literature, we present it here again in order to make the discussion of the comparison of the methodology with global analysis in Section 5 more concrete. Since it is known in advance that negative values in an elution profile or

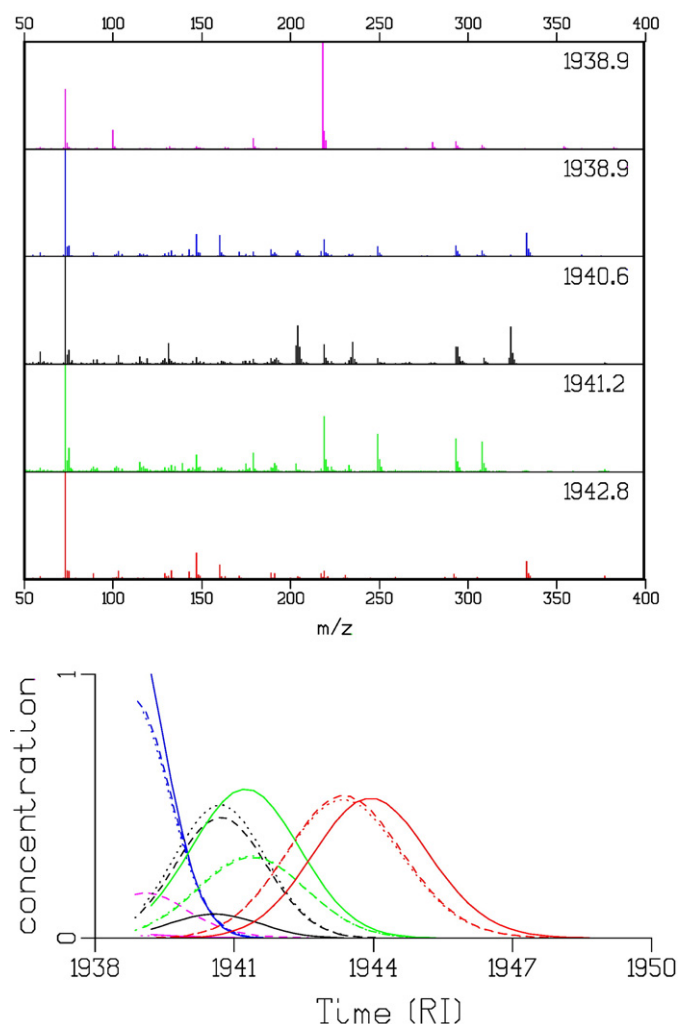


Fig. 10. Mixture II. (Top) Normalized mass spectra with location λ of peak maximum indicated in the upper right corner. (Bottom) Estimated elution profiles with colors indicating the five components, and linetypes indicating the different experiments. The contribution of each component to each data set is the product of each elution profile times the associated normalized mass spectrum times the scaling factor of 538,383. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Identifications for case studies Mixture I and Mixture II described in Section 3.2

Name	RI	MF	RMF	RI _{lit}
L-tyrosine	1938.9	795	882	1939
Galacturonic acid	1938.9	755	816	1940
Coniferylalcohol	1940.6	911	931	1944
Trans-p-coumaric acid	1941.2	887	900	1947
Galacturonic acid	1942.8	841	874	1961
Trans-caffeic acid (3TMS)	2133.1	949	949	2141

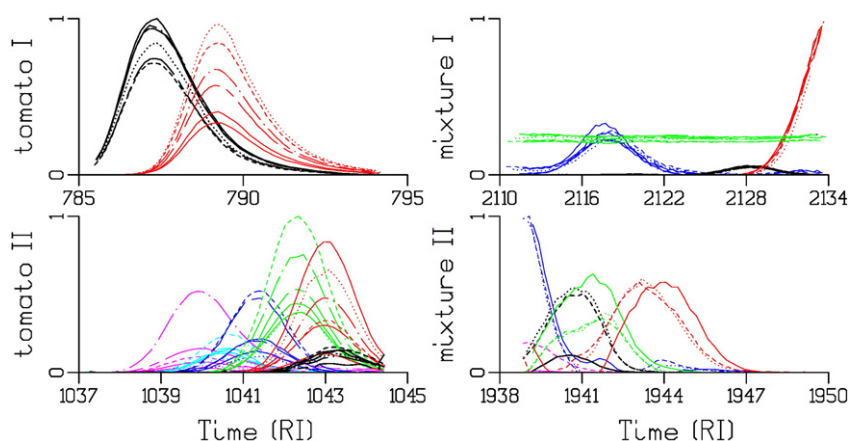


Fig. 11. Elution profiles estimated for the four case studies considered in Sections 3.1 and 3.2 after the refinement described in Section 2.6, with the scaling conventions as in Figs. 3 (Bottom), 6 (Bottom), 8 (Bottom), and 10 (Bottom).

mass spectrum are not physically interpretable, we present the basic algorithm with non-negativity constraints on both $C_1, \dots, C_K$ and S .

Algorithm 1 alternates calls to Algorithms 2 and 3, optimizing $C_1, \dots, C_K$ and S according to non-negative least squares criteria while considering either S or $C_1, \dots, C_K$ as fixed. The stopping criterion is usually based on the change in the residual difference between two iterations falling beneath some threshold. Note that 'rbind' refers to binding matrices together along their rows, and 'nrow' and 'ncol' refer to the number of rows and columns in a given matrix, respectively. 'NNLS' refers to a non-negative least squares algorithm such as that of Lawson and Hanson [25,26], a FORTRAN implementation of which is freely available via the Netlib repository [36] and in the R package **nnls** [37].

In practice, Algorithm 1 is often modified to include other constraints on $C_1, \dots, C_K$ and S , the most important of which are described in Section 4.1.

Algorithm 1. MCR-ALS($C_1, \dots, C_K, S, D_1, \dots, D_K$)

```

1:  $D_{All} := \text{rbind}(D_1, \dots, D_K)$ 
2: for  $i = 1, 2, \dots$ , until stopping criteria are met do
3:   if  $i$  is even then
4:      $S := \text{getS}(C_1, \dots, C_K, S, D_{All})$ 
5:   else
6:      $C_1, \dots, C_K := \text{getC}(C_1, \dots, C_K, S, D_1, \dots, D_K)$ 
7:   end if
8: end for
9: return  $C_1, \dots, C_K, S$ 

```

Algorithm 2. $\text{getS}(C_1, \dots, C_K, S, D_{All})$

```

1:  $C_{All} := \text{rbind}(C_1, \dots, C_K)$ 
2: for  $i = 1, 2, \dots, \text{ncol}(D_{All})$  do
3:    $S[i,] := \text{NNLS}(C_{All}, D_{All}[i,])$ 
4: end for
5: return  $S$ 

```

Algorithm 3. $\text{getC}(C_1, \dots, C_K, S, D_K)$

```

1: for  $j = 1, 2, \dots, K$  do
2:   for  $i = 1, 2, \dots, \text{nrow}(D_j)$  do
3:      $C_j[i,] := \text{NNLS}(S^T, D_j[i,])$ 
4:   end for
5: end for
6: return  $C_1, \dots, C_K$ 

```

Before MCR-ALS based on Algorithm 1 can be applied, it is necessary to decide what constraints to include, what number of

components (i.e., columns of C_j and S) to use, and what the starting estimates for either $C_1, \dots, C_K$ or S should be. These prerequisites are described in turn.

4.1. Constraints for use in MCR-ALS

While MCR-ALS works in the absence of a parametric model, it allows the application of constraints to the elution profiles and mass spectra. Commonly applied constraints include non-negativity of the elution profiles and mass spectra [38,39], unimodality of elution profiles [40], selectivity or equality constraints that account for intervals of data or datasets where a component is known to have zero amplitude or known amplitude [2], and normalization or closure constraints that help avoid problems associated with intensity ambiguities [41].

Unimodality constraints may be desirable in the case that the estimated elution profiles returned by Algorithm 1 have a bi-or-multi-modal shape that is believed to be unlikely given knowledge of the experimental set-up. For the application of unimodality constraints a simple modification of Algorithm 1 suffices; before returning an elution profile, the unimodal vector that is closest to the estimated elution profile in a least squares sense is determined and returned. The unimodal vector to be returned may be efficiently determined via a method based on isotonic regression [42] and implemented in the R package **Iso** [43].

In the case that it is known that a component is not present in a given dataset, a selectivity constraint may fix its elution profile to zero (likewise, if the component has a known elution profile, the profile may be fixed at the known shape). A heuristic that has been successfully applied to the analysis of GC/MS datasets checks if the retention time of an elution profile is within some threshold value of the median retention time for that component over all datasets. If not, the elution profile is set to zero (that is, its contribution is removed) in the deviating dataset [14].

Normalization and closure constraints, which are discussed in more detail elsewhere [41], may be of use to deal with the problem of intensity ambiguity. For any scalar m and elution profile c_i and mass spectrum s_i^T , $c_i m(1/m)s_i^T = c_i s_i^T$, meaning that increasing the estimate of c_i by a factor of m gives a model that fits the data equally well, as long as the mass spectrum s_i^T is multiplied by a factor of $1/m$. The normalization constraint is typically applied to the spectra, and may constrain $\|s_i^T\|_2 = 1$, or constrain the maximum of s_i^T to be one. A closure constraint is usually implemented on the rows $C_1, \dots, C_K$, and constrains the sum of the elements of each row of matrix C_j to be equal to a known constant, as for instance is desirable for reaction-based systems, where the elution profiles obey a mass balance

equation. Typically application of either normalization or closure is desirable, but not both types of constraint simultaneously.

4.2. Data selection and pre-processing

In the case that the mass spectrometry data contain on the order of 1–10 components, it is possible to treat all data at once. Otherwise, for the case of complex samples containing hundreds or thousands of different components, the resolution problem is approached by dividing the datasets into time intervals, so that data $D_1, \dots, D_K$ represent time windows of the full data that contain on the order of 1–10 components. Typically the time-windows are demarcated by points of low intensity, so that each set of data matrices $D_1, \dots, D_K$ contains the entire elution profile of the represented components [14].

In addition to time-window selection, baseline subtraction, alignment of datasets by peak matching and data smoothing are commonly performed prior to application of MCR-ALS. These pre-processing techniques are something of an art in themselves, and the optimal set of techniques is highly dependent on the particular experimental conditions [14].

4.3. Number of components

While many heuristics exist in the literature for the determination of the number of components present in a dataset D_i , methods based on principal component analysis (PCA) [44] or a singular value decomposition seem most popular [45], and are based on the assumption that any components to be resolved contribute more to the data than the noise term. For application to spectroscopy data, a heuristic has been suggested that performs MCR-ALS with an increasing number of components until the greatest number of components is found such that the retention times of the estimated elution profiles are in the same order in each dataset [14]. If it is known in advance that there is a baseline term in the data, it may be modeled with an extra component not subject to unimodality constraints [17].

4.4. Starting estimates

Most heuristics for obtaining starting values derive estimates for $C_1, \dots, C_K$ as opposed to S . Popular methods to obtain starting values for $C_1, \dots, C_K$ involve the use of evolving factor analysis [46] and the SIMPLISMA algorithm [14,47]. Algorithm 1 is sensitive to starting values in that both the final estimates for $C_1, \dots, C_K$ and S and the number of iterations required to meet stopping criteria depend on the initial values of $C_1, \dots, C_K$ or S .

4.5. Implementation

The MCR-ALS methods discussed are publicly available under the terms of the GNU General Public License in the form of the R packages **ALS** [48] from the Comprehensive R Archive Network (CRAN) [19]. The package includes scripts to repeat the MCR-ALS results described in Section 5.

5. Comparison of global analysis and MCR-ALS

MCR-ALS proceeds by iteratively solving constrained least squares problems, whereas global analysis requires the solution of a nonlinear optimization problem that usually takes the form of a separable nonlinear regression. For typical problems, MCR-ALS will be much faster than global analysis (requiring on the order of seconds as opposed to minutes). Furthermore, MCR-ALS requires significantly less hand-work in determining the model form.

Global analysis describes the data using a small number of free parameters as compared to MCR-ALS. For example, for the analysis of

two datasets where each dataset represents two components having elution profiles described by an EMG with the same shapes but different amplitudes per-dataset, only 8 intrinsically nonlinear parameters completely determine the model. In contrast, the number of free parameters in each iteration of MCR-ALS is equal to the number of entries in $C_1, \dots, C_K$ or S . Whether the large number of free parameters employed by MCR-ALS is an advantage or a liability depends on the specifics of the component resolution problem. For datasets in which the elution profiles are difficult to describe in terms of a model with a small number of parameters, MCR-ALS may provide a significantly better fit, due to its flexibility. However, as we will show in the remainder of this section, the relatively large number of parameters used by MCR-ALS may preclude the resolution of components in certain situations.

It is well-established that MCR-ALS is a useful and powerful component resolution tool. Therefore we will not concentrate on examples in which MCR-ALS is successful, which are abundant in the literature, but rather on those problems in which it performs poorly in terms of the matching factor of the estimated mass spectra as compared to global analysis. Many properties of the data affect the performance of both algorithms, such as the presence or absence of shape differences in the elution profiles, overlap/distance in time between locations of the elution profiles, amplitude differences of components between datasets, relative abundances of components, similarity of the underlying mass spectra, signal-to-noise ratio of the data, starting values, and number of datasets available for simultaneous analysis. The present study necessarily examines only a subset of the problem instances possible to encounter as these properties vary, though indicates some strengths and weaknesses of the algorithms that may be extrapolated to many other cases.

5.1. A basic example

We consider an example in which the data consists of two simulated GC/MS datasets that each represent two co-eluting components, shown in Fig. 12, with associated mass spectra shown in the top row of Fig. 13. This simulated data is inspired by the Mixture I case study described in Section 3.2. The parameters for the EMGs used to simulate the elution profiles are given in Table 3. The datasets represent integer time points 5720–5800 and integer masses 50–449.

In order to introduce a stochastic element into the data (that is, noise) we let a deterministic data point represent the parameter λ of a Poisson distribution. The Poisson distribution has density

$$p(x) = \lambda^x \exp(-\lambda) / x! \quad (10)$$

for $x=0, 1, 2, \dots$. The mean and variance are $E(X) = \text{Var}(X) = \lambda$. The signal-to-noise ratio (SNR) of a matrix of data with Poisson distributed noise is defined to be the square root of the maximum deterministic data point. To obtain data with SNR 1000 (which is encountered in GC/MS experiments) we scale the deterministic simulated dataset to contain 10^6 counts in the maximum data point, and then consider each data point to represent the λ in a Poisson process via the R function `rpois`.

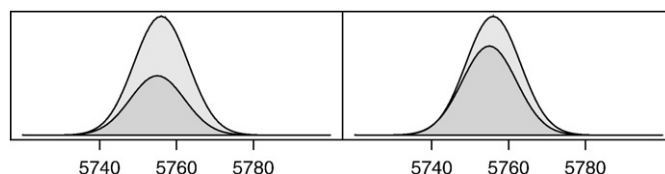


Fig. 12. The elution profiles represented in the two simulated datasets. In dataset 1 (left) the profiles have slightly different amplitude compared to the profiles in dataset 2 (right). Both components are represented by EMG functions, with the location parameter of the first component (dark grey) 1 time unit before that of the second component (light grey). Parameter values used to generate these profiles over the vector of integer times 5720–5800 are given in Table 3.

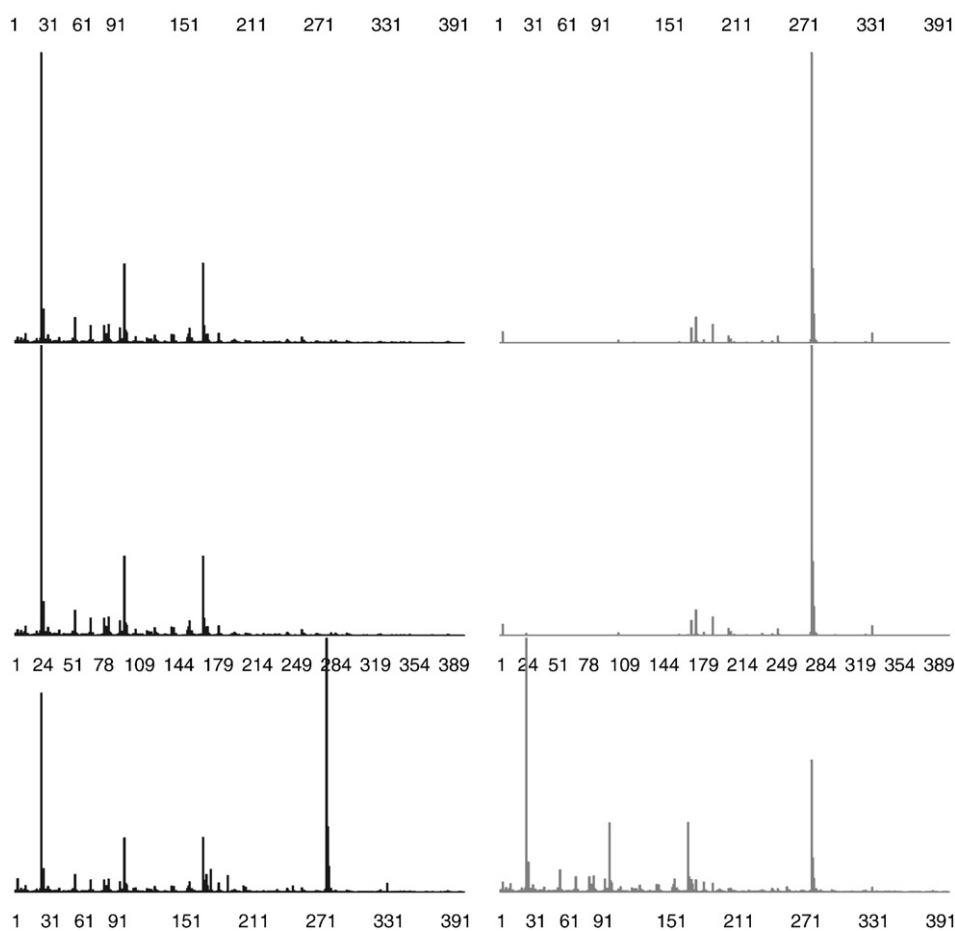


Fig. 13. The spectra on the top row are those used in simulating the data. The second row contains representative spectra estimated by global analysis. The third row contains representative spectra estimated by MCR-ALS, in which the spectra are not well resolved.

In order to judge how well an estimated mass spectrum s resembles the mass spectrum u used in simulating the data, the normalized dot product matching factor function

$$\cos(u, s) = \frac{u \cdot s}{\|u\| \|s\|} \quad (11)$$

is useful [18,49]. The normalized dot product matching factor of the spectra shown in the top row of Fig. 13 is almost 0, as the spectra are very dissimilar. If the abundance ratio of components p and q is defined as

$$ab(p, q) = \frac{\sum(C_{j[, p]} S_{[, p]}^T)}{\sum(C_{j[, q]} S_{[, q]}^T)} \quad (12)$$

where $C_{j[, p]}$ and $C_{j[, q]}$ are the elution profiles p and q over all times, $S_{[, p]}$ and $S_{[, q]}$ are the mass spectra p and q over all masses, and the sum function takes the sum of all matrix elements. In this way Eq. (12) represents the number of counts contributed to the data by the component p divided by the number of counts contributed by component q . Under this definition the abundance ratio of component 1 to component 2 is ≈ 2.1 in dataset 1 and ≈ 3.1 in dataset 2; the difference in abundance ratios between datasets is due to the difference in the amplitude of the first elution profile.

In applying MCR-ALS to resolving components in the simulated data, we apply constraints for non-negativity of $C_1, \dots, C_K$ and S , unimodality of $C_1, \dots, C_K$, and normalization of each mass spectrum such that the maximum value is one. For global analysis, the mass spectra and amplitudes of the elution profiles are constrained to non-

negative values. The starting values for both MCR-ALS and global analysis are taken to be the elution profiles used in simulation, but with location parameters shifted to 5757 and 5753. The stopping criterion used for MCR-ALS is reduction of the residual difference by no more than 0.001 between iterations or completion of 100 iterations. The stopping criteria for global analysis were the defaults used by the Levenberg–Marquardt non-linear regression algorithm implemented in the R package **minpack.lm**, or completion of 50 iterations. In all problems considered in this study, allowing the algorithms to run for more iterations results in further reductions in the sum of squared errors but does not change the matching factor of the estimated mass spectra with the mass spectra used in simulation by more than $\approx 5\%$.

Fig. 13 shows that for this problem, MCR-ALS does not resolve the spectral signature of both components, whereas global analysis estimates the mass spectra well. The normalized dot product matching factor of the estimated spectra and the spectra used in simulation over 100 noise realizations of the problem were calculated. Global analysis results in matching factors for both components that are always > 0.99 , whereas MCR-ALS does not well-resolve both

Table 3

Tabulated are the parameters for the two EMGs used to describe the elution profiles in both datasets

	Component 1	Component 2
Location	5754	5755
FWHM	7	7
Rate	1	1

The elution profiles have amplitudes 1 and 2, respectively, in dataset 1, and amplitudes 1.5 and 2 respectively, in dataset 2. The resulting EMGs are shown graphically in Fig. 12.

Table 4
Eleven pairs of datasets are simulated for the studies in Sections 5.2 and 5.3, each of which represents elution profiles over integer times 5720–5800 with different distances between the location of the components

Dataset pair	1	2	3	4	5	6	7	8	9	10	11
Location separation	0.01	0.05	0.1	0.25	0.5	1	2	3	4	5	6

The resulting EMGs are shown graphically in Fig. 14.

spectra. By examination of many stochastic realizations of the data, it is clear that the results of both algorithms are stable with respect to this level of noise. We turn to a more thorough simulation study to further investigate the resolution power of the techniques.

5.2. Varying distance between components

The example put forth in Section 5.1 can be studied as the location of component 1 with respect to the location of component 2 is made to vary. We simulate pairs of datasets with different distances between the locations of the components, as collated in Table 4 and shown graphically in Fig. 14. All other experimental parameters are as described previously in Section 5.1.

For each pair of datasets, 25 different stochastic realizations are generated. The left plots of Fig. 15 show the average matching factor of the mass spectra estimated by MCR-ALS and global analysis with the mass spectra used in simulation over 25 stochastic realizations of each pair of datasets. As the distance between the components increases, MCR-ALS is better able to solve the problem, but does not resolve components when the distance between the locations of the components is small. Global analysis is able to estimate both mass spectra well in all cases. When the distance between the components is reduced to zero, both global analysis and MCR-ALS do not resolve the underlying mass spectra. Both methods also fail in the case that only one of the two datasets is analyzed.

5.3. Mass spectra with increased matching factor

We have performed the same simulation study as in Section 5.2 but using mass spectra with a larger matching factor (0.31 as opposed to 0), inspired by the Tomato I case study described in Section 3.1, which are shown in Fig. 3 (Top). We again simulate pairs of datasets with varying locations of the first component, as collated in Table 4 and shown graphically in Fig. 14. The right plots in Fig. 15 show the average matching factor of the mass spectra estimated by MCR-ALS and global analysis with the mass spectra used in simulation, again as averages over 25 stochastic realizations of each pair of datasets. As in Section 5.2, global analysis is better able to estimate the underlying mass spectra when the separation between the location of the components is small.

5.4. Varying SNR

Returning to the study in Section 5.2, we choose a distance between the components for which both global analysis and MCR-ALS estimated the mass spectra of both components well, namely the problem instance with elution profiles separated by 6 time units described in Section 5.2, and a distance for which global analysis succeeds in estimating the mass spectra well, but MCR-ALS does not, namely the problem instance with the elution profiles separated by 1 time unit described in Section 5.2.

We simulated these problem instances for SNR 500, 200, 100 and 25, to examine the robustness of the results to noise, examining the performance of the algorithm in terms of the matching factor of the estimated spectra with the mass spectra used in simulating the data. For all these SNRs, for both problems, the algorithms performed qualitatively the same as under SNR 1000, with differences of less than <5% in the resulting average matching factors of the estimated spectra compared to the values used in simulation over 25 noise realizations

for each SNR considered. Hence the results described in previous sections are stable with respect to the noise level of the data.

5.5. Sensitivity to starting values and stopping criteria

For the problem considered in Section 5.1, using starting values for the elution profiles that are slightly shifted (e.g., by three time units) from the values used to simulate the elution profiles, MCR-ALS continues to reduce the sum of squared errors for many iterations (>10,000) provided the stopping criterion is altered to allow continuation even if the differences in the residuals between iterations are very small. Eventually MCR-ALS reaches the same minimum in sum of squared errors it finds in only a few iterations when given perfect starting values for the elution profiles. However, the estimated

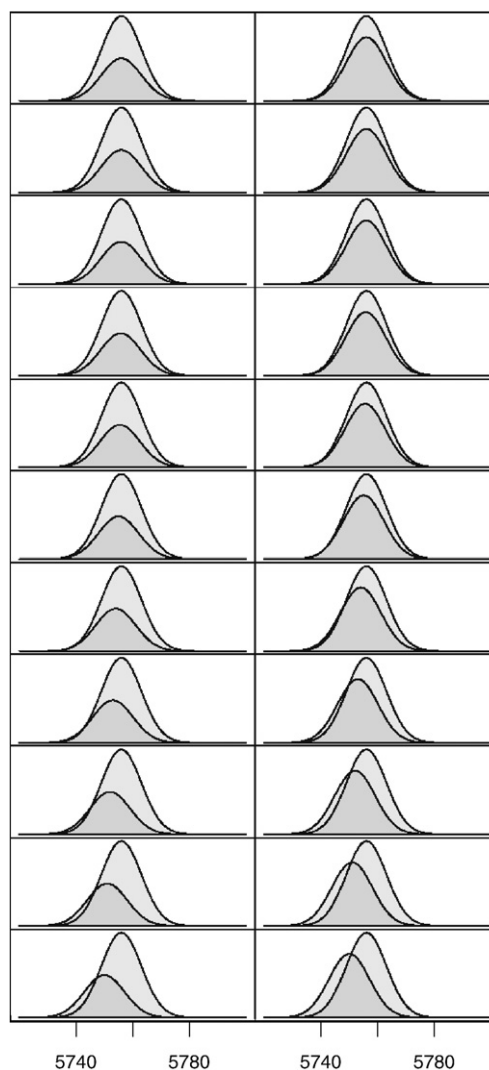


Fig. 14. Elution profiles used in simulating the pairs of datasets used in Sections 5.2 and 5.3; each row represents a pair of datasets with a different location of the first component. Each elution profile is represented by an EMG having the parameter values given in Table 3, except for the location of the first component, which is less than the location of the second component by the values given in Table 4.

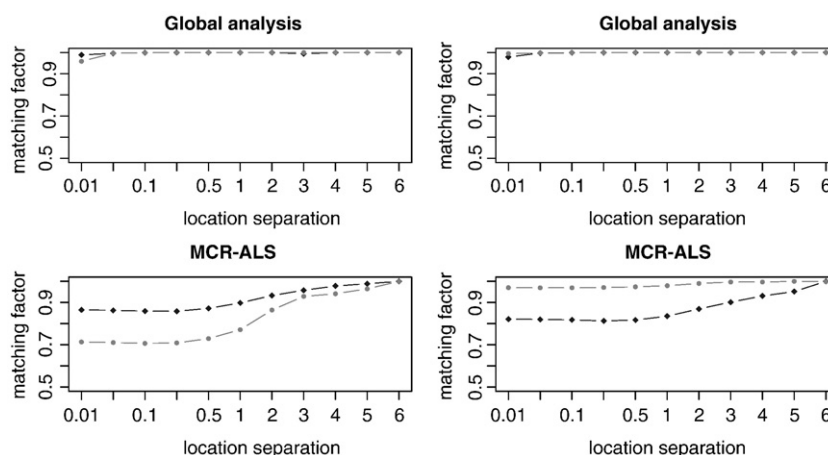


Fig. 15. The normalized dot product matching factor of the estimated spectra and the spectra used in simulation for global analysis (top) and MCR-ALS (bottom). Results for component 1 are shown as black diamonds, and for component 2 as grey circles. As the separation of the components is increased, the component resolution problem is rendered progressively easier. The left plots are for the data in Section 5.2, whereas the right plots are for the data in Section 5.3. MCR-ALS well-resolves the spectra only at a separation of 6.

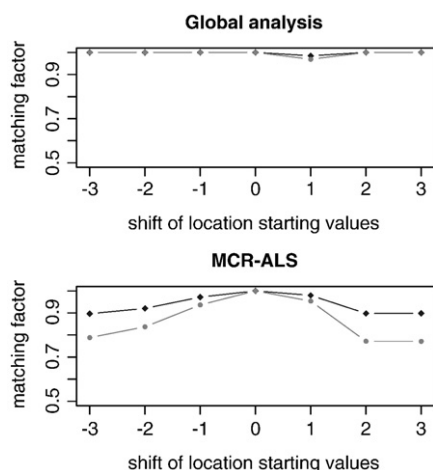


Fig. 16. The normalized dot product matching factor of the estimated spectra and the spectra used in simulation for global analysis (top) and MCR-ALS (bottom) as the starting value for the location of the elution profiles is shifted in time from the values used in simulation. Results for component 1 are shown as black diamonds, and for component 2 as grey circles. MCR-ALS converges to mass spectra estimates that are different from the actual underlying mass spectra when the starting values are less than perfect. The results shown are average matching factors over 10 stochastic realizations of the data.

spectra do not well-represent the spectra used in simulation, and are associated with a matching factor of <0.9 . There are therefore a range of values for the elution profiles and mass spectra that result in the same sum of squared errors under MCR-ALS. That is, there are bands of feasible solutions, not a unique description of the elution profiles and mass spectra that result in the best model fit. This is due to the relatively large number of parameters used for the description of the elution profiles under MCR-ALS, and is an issue that has been studied in detail elsewhere [41,50]. Fig. 16 shows that shifting the starting values for the elution profiles from the values used in simulation for the problem described in Section 5.1 results in varying estimates for the mass spectra under MCR-ALS. This is in contrast to the estimates for the mass spectra returned by global analysis, which always well approximate the mass spectra used in simulation.

5.6. Sensitivity to abundance ratio of components

By increasing the magnitude of the first mass spectrum relative to the magnitude of the second mass spectrum used in simulating the data, it is possible to vary the number of counts contributed to the data by each of the two components, so that the abundance ratio defined in Eq. (12) varies. For the problem described in Section 5.2 where the distance between the location of the components is 6 time units, and both MCR-ALS and global analysis estimate the components well, we

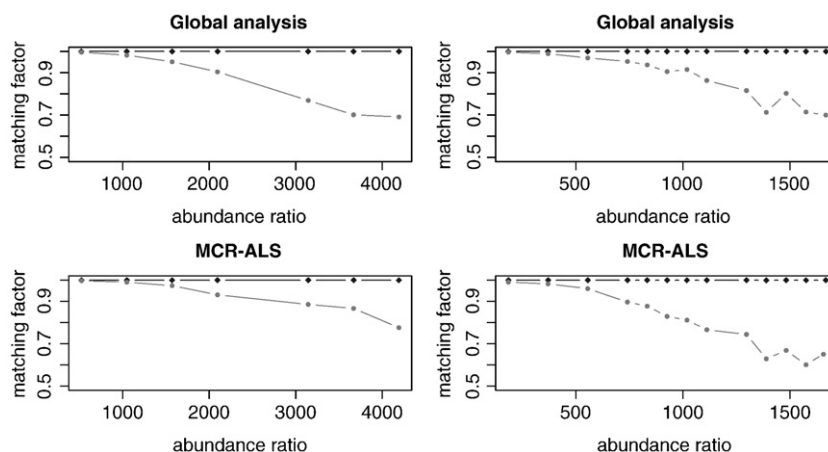


Fig. 17. The normalized dot product matching factor of the estimated spectra and the spectra used in simulation for global analysis (top plots) and MCR-ALS (bottom plots) as the abundance of the first component relative to the second is made to increase. Matching factors shown are averages over 25 stochastic realizations of the data. The left and right plots are for data with the underlying mass spectra shown in the top row of Fig. 13 and in Fig. 3 (Top), respectively.

examined how the performance varies in terms of the estimated mass spectra as the abundance ratio of component 1 to component 2 is made to vary. We also examined this same problem but using the spectra shown in Fig. 3 (Top). The breakdown of the resolution power of the techniques was qualitatively similar as the abundance of component 1 to component 2 increases, as shown in Fig. 17.

6. Discussion

While global analysis is a powerful component resolution tool, much additional work is necessary to render it a suitable tool for high-throughput applications. Of particular importance is the development of heuristics for model specification.

Other tools for component resolution of mass spectrometry data such as the “Automated Mass Spectral Deconvolution and Identification System” (AMDIS) by the National Institute of Standards and Technology [34] have the great advantage of being much more fully automatic, and are integrated with libraries of mass spectra so as to best facilitate the end goal of compound identification. However, they are not able to deal with the problem of completely overlapping components or multiple datasets. The ideal software would combine the best features of all available methodologies. Then when one of the faster methods based on peak-finding in the chromatogram or MCR-ALS fail to resolve components, global analysis could be applied. This would require the formulation of heuristics to decide when application of global analysis is likely to offer improved estimates of the underlying spectra. Possible heuristics could be based on rank analysis of the data, or prior knowledge indicating the expected number of co-eluting components present.

Further research into the utility of methods suggested by [12,13], that are hybrid in the sense that they use MCR-ALS to iteratively improve estimates for $C_1, \dots, C_K$ and S , but refine the MCR-ALS estimates for $C_1, \dots, C_K$ by fitting these elution profiles with a parametric model, is also of interest for the problems discussed here. Such methods are promising for reducing the sensitivity to starting values of MCR-ALS studied in Section 5.5.

7. Conclusions

Benefits of global analysis are: (1) direct quantitative estimation of parameters of interest, and (2) the quality of the fit can be judged from residual analysis. The parameters of interest are the mass spectra of the components, and their elution profiles, in particular the Retention Index that can be calculated from the location parameter of each elution profile. The precision of these parameters can be estimated as well. If the singular vectors of the residual matrix show only little structure, this indicates that the data have been fitted up to the noise level, and thus all information has been extracted. If they do show structure, this can indicate that the assumed EMG shape is not fully adequate, or it can indicate that an extra component is needed to fit the data. These benefits apply when analyzing single or multiple datasets. However, with simultaneous global analysis of multiple datasets there are more benefits. Overlapping components can more easily be resolved when they are present in different amounts, which occurs naturally when comparing biological samples. The estimated parameters (in particular the mass spectra) are more robust against systematic measurement errors. The nuisance parameters (most importantly for alignment) can be directly estimated as well. The model allows the imposition of common shapes for elution profiles across data sets, and common retention time differences between components across data sets.

Global analysis and MCR-ALS methodologies for component resolution in time-resolved mass spectrometry data have been compared in simulation studies in which the elution profiles used to generate the data are completely overlapping. Problem instances have been presented in which global analysis is able to estimate the mass

spectra of all components well, whereas MCR-ALS fails to do so. Such problem instances arise when elution profiles are completely overlapping, making the component resolution task difficult.

While global analysis is powerful for component resolution, especially for data representing co-eluting compounds, it requires more run-time, and at present is less automated than other techniques based on peak-finding and MCR-ALS. Possible improvements in accuracy offered by the method must therefore be carefully weighed against losses in efficiency. Further study and development of the method for mass spectrometry applications is warranted.

Acknowledgments

Yury Tikunov and Harrie Verhoeven of Plant Research International kindly provided the tomato data discussed in Section 3.1. Thomas Moritz, Pär Jonsson and Krister Lundgren from Umeå Plant Science Center generously provided the standard mixture data described in Section 3.2. Francel Verstappen, Egon Willighagen, Arjen Lommen, Hans van Beek, Thomas Binsl, Yury Tikunov and Harrie Verhoeven are thanked for helpful discussions. KMM was supported by Computational Science grant 635.000.014 from the Netherlands Organization for Scientific Research (NWO). VVM was supported by the Netherlands Bioinformatics Centre (NBIC).

References

- [1] J. Jiang, Y. Liang, Y. Ozaki, *Chemometrics and Intelligent Laboratory Systems* 71 (2004) 1–12.
- [2] R. Tauler, A. Smilde, B. Kowalski, *Journal of Chemometrics* 9 (1995) 31–58.
- [3] R. Tauler, *Chemometrics and Intelligent Laboratory Systems* 30 (1995) 133–146.
- [4] A. de Juan, R. Tauler, *Analytica Chimica Acta* 500 (2003) 195–210.
- [5] M. Garrido, F.X. Rius, M.S. Larrechi, *Analytical and Bioanalytical Chemistry* 390 (2008) 2059–2066.
- [6] G.H. Golub, R.J. LeVeque, *Proceedings of the 1979 Army Numerical Analysis and Computers Conference*, 1979, pp. 1–12.
- [7] J.F. Nagle, *Biophysical Journal* 59 (1991) 476–487.
- [8] I.H.M. van Stokkum, D.S. Larsen, R. van Grondelle, *Biochimica et Biophysica Acta* 1657 (2004) 82–104 and erratum, 1658, 262.
- [9] S. Pelet, M.J.R. Prevorse, L.H. Laiho, P.T.C. So, *Biophysical Journal* 87 (2000) 2807–2817.
- [10] P.J. Verwee, A. Squire, P.H. Bastiaens, *Biophysical Journal* 78 (2000) 2127–2137.
- [11] S. Liptonok, K.M. Mullen, J.W. Borst, I.H.M. van Stokkum, V.V. Apanasovich, A.J.W.G. Visser, *Journal of Statistical Software* 18 (8) (2007) 1–20. <http://www.jstatsoft.org/v18/i08/>.
- [12] A. de Juan, M. Maeder, M. Martínez, R. Tauler, *Analytica Chimica Acta* 442 (2001) 337–350.
- [13] A. de Juan, M. Maeder, M. Martínez, R. Tauler, *Chemometrics and Intelligent Laboratory Systems* 54 (2000) 123–141.
- [14] P. Jonsson, A. Johansson, J. Gullberg, J. Trygg, J.A.B. Grung, S. Marklund, M. Sjöström, H. Antti, T. Moritz, *Analytical Chemistry* 77 (2005) 5635–5642.
- [15] E. Peré-Trepát, S. Lacorte, R. Tauler, *Analytica Chimica Acta* 595 (2007) 228–237.
- [16] S. Mas, G. Fonrodona, R. Tauler, J. Barbosa, *Talanta* 71 (2007) 1455–1463.
- [17] E. Peré-Trepát, S. Lacorte, R. Tauler, *Journal of Chromatography A* 1096 (2005) 111–122.
- [18] S.E. Stein, D.R. Scott, *Journal of the American Society for Mass Spectrometry* 5 (1994) 859–866.
- [19] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria 3-900051-07-0, 2008.
- [20] G.H. Golub, V. Pereyra, *Inverse Problems* 19 (2003) R1–R26.
- [21] M. Maus, M. Cotlet, J. Hofkens, T. Gensch, F.C. de Schryver, J. Schaffer, C.A.M. Seidel, *Analytical Chemistry* 73 (2001) 2078–2086.
- [22] V.B.D. Marco, G.G. Bombi, *Journal of Chromatography A* 931 (2001) 1–30.
- [23] D.M. Sima, S. Van Huffel, *Journal of Computational and Applied Mathematics* 203 (2007) 264–278.
- [24] K.M. Mullen, *Separable nonlinear models: theory, implementation and applications in physics and chemistry*, Ph.D. thesis, Department of Physics and Astronomy, Vrije Universiteit Amsterdam, The Netherlands (2008).
- [25] C.L. Lawson, R.J. Hanson, *Solving Least Squares Problems*, Prentice Hall, Englewood Cliffs, NJ, 1974.
- [26] C.L. Lawson, R.J. Hanson, *Solving Least Squares Problems*, Classics in Applied Mathematics, SIAM, Philadelphia, 1995.
- [27] J.W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, Reading, Massachusetts, 1977.
- [28] Y. Tikunov, A. Lommen, C.R. de Vos, H.A. Verhoeven, R.J. Bino, R.D. Hall, A.G. Bovy, *Plant Physiology* 139 (2005) 1125–1137.
- [29] K.M. Mullen, I.H.M. van Stokkum, *Journal of Statistical Software* 18 (3) (2007) 1–46. <http://www.jstatsoft.org/v18/i03/>.
- [30] P. Ausloos, C. Clifton, S. Lias, A. Mikaya, S. Stein, D. Tchekhovskoi, O. Sparkman, V. Zaikin, D. Zhu, *Journal of the American Society for Mass Spectrometry* 10 (1999) 287–299.

- [31] J. Kopka, N. Schauer, S. Krueger, C. Birkemeyer, B. Usadel, E. Bergmuller, P. Dormann, W. Weckwerth, Y. Gibon, M. Stitt, L. Willmitzer, A.R. Fernie, D. Steinhauser, *Bioinformatics* 21 (2005) 1635–1638.
- [32] V.V. Mihaleva, H.A. Verhoeven, R.C. de Vos, R.D. Hall, R.C. van Ham, submitted for publication.
- [33] M. Petro-Turza, *Food Reviews International* 2 (1987) 309–351.
- [34] S.E. Stein, *Journal of the American Society for Mass Spectrometry* 10 (1999) 770–781.
- [35] R. Buttery, G. Takeoka, *Journal of Agricultural and Food Chemistry* 52 (2004) 6264–6266.
- [36] S.V. Browne, J.J. Dongarra, S.C. Green, K. Moore, T.H. Rowan, R.C. Wade, Netlib services and resources, Technical Report CS-94-222, University of Tennessee, 1994.
- [37] K.M. Mullen, I.H.M. van Stokkum, **nnls**: The Lawson-Hanson algorithm for non-negative least squares (NNLS) (2007), R package version 1.2.
- [38] R. Bro, S.D. Jong, *Journal of Chemometrics* 11 (1997) 393–401.
- [39] M.H.V. Benthem, M.R. Keenan, *Journal of Chemometrics* 18 (2004) 441–450.
- [40] R. Bro, N.D. Sidiropoulos, *Journal of Chemometrics* 12 (1998) 223–247.
- [41] R. Tauler, *Journal of Chemometrics* 15 (2001) 627–646.
- [42] T.R. Turner, P.C. Wollan, *Computational Statistics & Data Analysis* 25 (1997) 305–320.
- [43] R. Turner, **Iso**: Functions to perform isotonic regression (2005), R package version 0.0-2.
- [44] S. Wold, K. Esbensen, P. Geladi, *Chemometrics and Intelligent Laboratory Systems* 2 (1987) 37–52.
- [45] G.H. Golub, C.F. van Loan, *Matrix Computations*, 3rd edn. The Johns Hopkins University Press, Baltimore, MD, USA, 1996.
- [46] M. Maeder, *Analytical Chemistry* 59 (1987) 527–530.
- [47] W. Windig, J. Guilment, *Analytical Chemistry* 63 (1991) 1425–1432.
- [48] K.M. Mullen, I.H.M. van Stokkum, **ALS**: multivariate curve resolution alternating least squares (MCR-ALS) (2008), R package version 0.0.2.
- [49] Z.B. Alfassi, *Journal of the American Society for Mass Spectrometry* 15 (2004) 385–387.
- [50] M. Garrido, M. Larrechi, F. Rius, R. Tauler, *Chemometrics and Intelligent Laboratory Systems* 76 (2005) 111–120.