On the analysis of chromatographic biopharmaceutical data by curve resolution techniques in the framework of the area of feasible solutions.

Mathias Sawall[a], Matthias Rüdt[b], Jürgen Hubbuch[b], Klaus Neymeyr[a,c]

[a] *Universität Rostock, Institut für Mathematik, Ulmenstraße 69, 18057 Rostock, Germany*
[b] *Karlsruhe Institute of Technology, Institute of Engineering in Life Sciences, Fritz-Haber-Weg 2, 76131 Karlsruhe, Germany*
[c] *Leibniz-Institut für Katalyse, Albert-Einstein-Straße 29a, 18059 Rostock, Germany*

Abstract

Monitoring preparative protein chromatographic steps by in-line spectroscopic tools or fraction analytics results in medium or large sized data matrices. Multivariate Curve Resolution (MCR) serve to compute or to estimate the concentration values of the pure components only from these data matrices. However, MCR methods often suffer from an inherent solution ambiguity which underlies the factorization problem. The typical unimodality of the chromatographic profiles of pure components can support the chemometric analysis. Here we present the pure components estimation process within the framework of the area of feasible solutions, which is a systematic approach to represent the range of all possible solutions. The unimodality constraint in combination with Pareto optimization is shown to be an effective method for the pure component calculation. Applications are presented for chromatograms on a model protein mixture containing ribonuclease A, cytochrome c and lysozyme and on a two-dimensional chromatographic separation of a monoclonal antibody from its aggregate species. The root mean squared errors of the first case study are 0.0373, 0.0529 and 0.0380 g/L compared to traditional off-line analytics. The second case study illustrates the potential of recovering hidden components with MCR from off-line reference analytics.

Key words: pure component calculation, multivariate curve resolution, bio-pharmaceuticals, chromatography, unimodality, FACPACK.

1. Introduction

Chromatography in biopharmaceutical industry is of great importance not only as an analytical technique but also as a preparation method for pure materials [1, 2, 3]. A typical application is protein chromatography [1, 3]. The basis for protein fractionation are different retention times of proteins due to their specific elution behavior, but the resulting chromatograms typically do not achieve baseline separation of different species. Instead, a second, orthogonal analytical measurement method is necessary to interpret the preparative protein chromatograms [1, 4, 5, 6]. The analytical dimension is typically achieved by off-line reference analytics. Alternatively, selective quantification by spectroscopic tools has been investigated as potential method for process monitoring [7]. The analysis of measurements is often done manually and is by no means trivial. Typically, the results depend on the analyst interpreting the data. In-line spectroscopic measurements are analyzed with calibrated statistical models such as partial-least squares regression [7]. Here, significant resources have to be invested into model calibration. The quality of the analysis has a distinct impact on the reliability and reproducibility of the results from preparative protein chromatography.

Automated calibration-free tools for chromatographic data analysis can improve the data interpretation and can support the reproducibility of the results [8]. Such tools should be capable to identify the number of species, to compute pure component factorizations, to exploit the existence of unimodal profiles within the decomposition process and, last but not least, to handle strongly interfering peaks. Ideally, the method should be robust with respect to variations of the observed retention times.

The solution of such data analytical problems is incumbent on multivariate curve resolution (MCR) methods [9, 10], which are effective tools for the analysis of spectroscopic and chromatographic data. MCR methods solve the pure component calculation problem, namely to reconstruct the (pure component) factors from mixed spectral or

Email: mathias.sawall@uni-rostock.de (Mathias Sawall), mruedt@icloud.com (Matthias Rüdt), juergen.hubbuch@kit.edu (Jürgen Hubbuch), klaus.neymeyr@uni-rostock.de (Klaus Neymeyr)

chromatographic data. Originally, MCR methods were developed to decompose spectral mixture data as reaction-accompanying series of IR/Raman or UV/Vis spectra into the contributions from the pure components [11, 12]. The application of MCR methods to chromatographic data to biopharmaceutical applications is comparatively new [13, 14, 15, 16].

Typically the focus of MCR-methods is on determining one, single factorization for the given (chromatographic) data, but no systematic analysis of the solution ambiguity is applied [13, 15, 17, 16]. It is worth stressing that MCR-methods by using constraint-based Pareto optimization techniques evade the issue of non-unique solutions [18, 19]. Consequently, the pure component factorizations as computed by the routines [20, 21] and others strongly depend on the active constraints and their relative weighting. Such a steering effect of validated constraints is very helpful. For chromatographic data the typical unimodal profile shape of single components can support the MCR analysis [22, 23]. The situation is more complex in the case of multiple components with overlapping profiles. Many solutions can exist, but most of them do not represent chemically and physically meaningful profiles.

The intention of this paper is to discuss the pure component factorization problem for chromatographic biopharmaceutical data on the basis of the so-called area of feasible solutions (AFS), [24, 25, 26, 27, 28, 29]. This permits a general approach for the analysis of the solution ambiguity of the pure component calculation problem. The application of this relatively young tool to chromatographic data and biopharmaceutical problems seems to be new. Our focus lies on analyzing the non-uniqueness of the solutions, on giving guidelines for the practical application of the methods and on interpreting its results. The AFS is a low-dimensional representation of all possible profiles that the factorization problem can have for a given data set. Being aware on the existence of a solution ambiguity can only be a first step. A second step should be a reduction to a single, hopefully true and chemically correct decomposition. We show that for the given chromatographic data the unimodality constraint can serve as an effective reducer in order to find the desired solution. This work takes up the chemometric MCR-based analyses of [14, 16] on the pure component extraction of preparative protein chromatograms and demonstrates the application of AFS-based chemometric analyses. We hope that the latter more general approach can provide a deeper insight into the problem of profile ambiguities and strategies how to select the correct solution.

1.1. Organisation of the paper

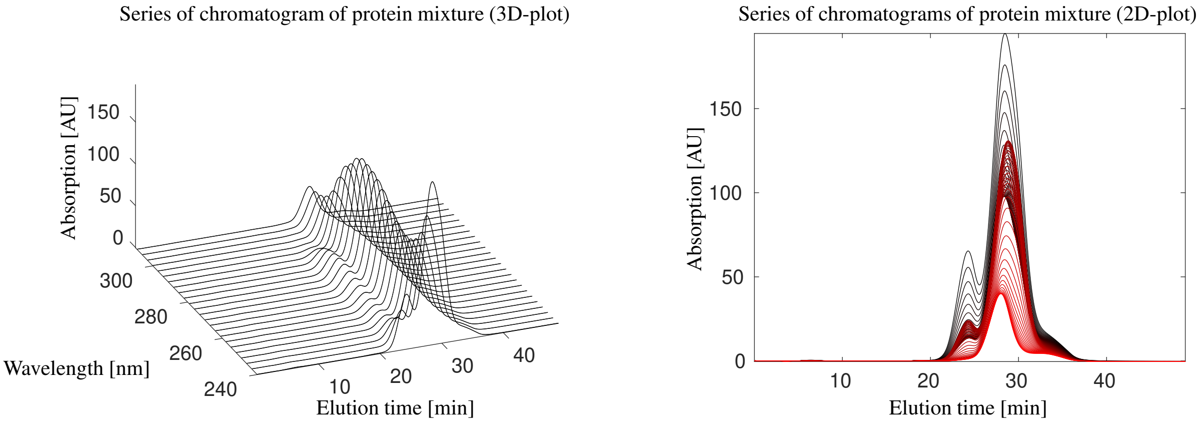
The paper is organized as follows. First, Section 2 introduces two experimental data sets for the subsequent method development and the numerical testing. Section 3 is of central importance. It introduces the theoretical background of the solution ambiguity underlying MCR methods and how it complicates the pure component calculation problem. The computation of properly constrained factors with respect to penalty and regularization functions is explained in Sec. 3.3. Then Sec. 3.4 presents the AFS approach and the FACPACK software for AFS related computations. Finally, Sec. 4 reports on the results for the experimental data sets and discusses the questions concerning the identification of the number of absorbing and reconstructable components.

2. Materials & Methods

The following two experimental data sets accompany the method developments and numerical experiments:

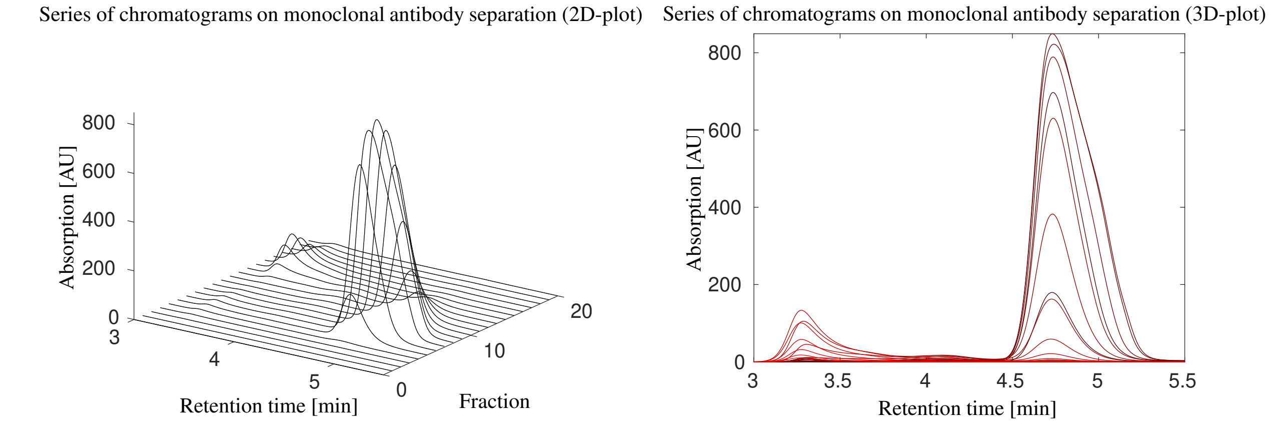
Data set 1. The first mixture contains the three model proteins, ribonuclease A, cytochrome c and lysozyme. The detailed data was previously published in [16]. Briefly, the UV/Vis-absorbance data of a cation-exchange chromatography step from 240 nm to 310 nm was used for this study. A number of $k = 588$ spectra in the interval $t \in [1.67 \cdot 10^{-2}, 48.93]$ min were taken. Each spectrum consists of absorption values at $n = 71$ wavelengths. The data (in 2D- and 3D-plots) as well as the singular value decomposition of the data matrix are shown in Figs. 1 and 3. (All numerical computations in this paper are done in MatLab version R2018a and all singular value decompositions are computed by its `svd`-routine.) MCR-based pure component factors under the constraint of unimodal elution profiles are shown in Fig. 5.

Data set 2. This data set describes a two-dimensional chromatographic separation of a monoclonal antibody (mAb) from its aggregate species. For technical details on this data set see [30]. The first chromatographic separation was based on preparative cation exchange chromatography (CEX) which is typically used in biopharmaceutical production processes for the purification of proteins [1]. However, preparative CEX only provides limited resolution of the different species. In order to specify the elution profiles, the effluent from the 5 column volume preparative CEX step was collected in fractions. These were then analyzed by a higher resolution second chromatographic separation (analytical size exclusion chromatography) to improve the resolution of the different species. During the second


Series of chromatogram of protein mixture (3D-plot)

Series of chromatograms of protein mixture (2D-plot)

Figure 1: Series of chromatograms of a protein mixture according to data set 1 in a 3D-plot (left, only every third wavelength) and a 2D-plot (right, color changing from red to black with increasing wavelength values).


Series of chromatograms on monoclonal antibody separation (2D-plot)

Series of chromatograms on monoclonal antibody separation (3D-plot)

Figure 2: Series of chromatograms of the separation of a monoclonal antibody (mAb) from its aggregate species according to data set 2. Left: Series of chromatograms for the 20 fractions. Throughout this paper absorption values are given in AU for Arbitrary Unit. Right: Two-dimensional plot of the chromatographic profiles with a color changing from black to red with increasing fraction values.

chromatographic separation, the UV absorbance trace at 280 nm was recorded. For the analysis we consider a number of $k = 21$ chromatograms each with absorption values at $n = 3751$ retention times in the interval $t \in [3, 5.5]$ min. The resulting series of analytical chromatograms are shown in two- and three-dimensional representation in Fig. 2. Figure 4 shows the dominant singular values and the associated singular vectors. The resulting pure component estimation is presented in Fig. 9.

3. Theory: The pure component calculation problem, its solution and solution ambiguties

The pure component calculation problem deals with the extraction of the pure component information (here chromatographic elution profiles) from the spectral or chromatographic mixture data. The series of spectra or chromatograms is stored row-wise in a matrix $D \in \mathbb{R}^{k\times n}$. These data correspond to measurements on a retention time×wavelength grid or a retention time×fraction grid. The fraction axis can be considered as an intermediate coordinate for the second chromatographic separation. The naming of this coordinate is not important for the following analysis. Instead, we are interested in factorising the matrix D of mixture data to its source terms $C \in \mathbb{R}^{k\times s}$ and $S \in \mathbb{R}^{n\times s}$ according the underlying bilinear structure of the problem. The dimension s equals the number of components. This bilinear structure is expressed in the well-known Lambert-Beer law reading in matrix notation

$$D = CS^T + E. \tag{1}$$

Ideally, the matrix E is the null matrix. In practice, E should be close to the null matrix and its non-zero entries express deviations from strict bilinearity, e.g., due to noise. The columns of C are the concentration profiles in time of the pure components. Depending on the problem, the columns of S are elution profiles, pure component spectra or other pure component profiles.

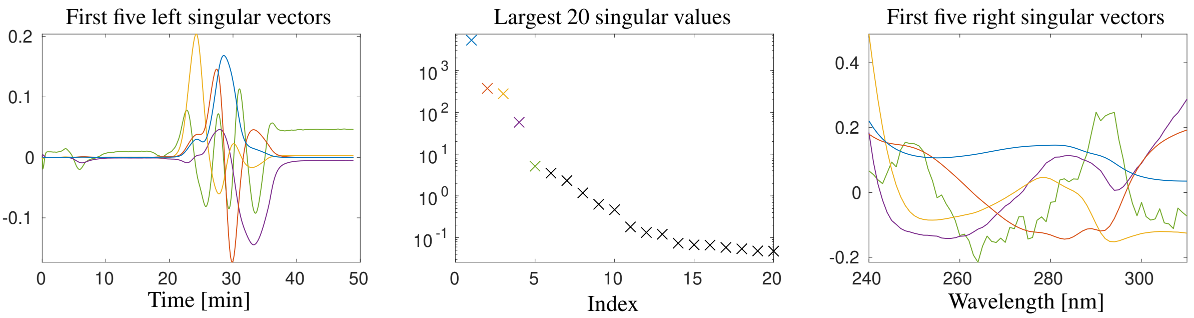


Figure 3: The dominant singular values and associated singular vectors by an SVD of the matrix underlying data set 1. The largest/dominant singular value and the associated singular vectors are drawn blue. Then with a decreasing size of the singular values the colors red, ochre, purple and green are used. The singular values indicate that at least $s = 3$, but not more than $s = 4$ components can be reconstructed.

3.1. Pure component calculation and the non-uniqueness of solutions

The problem of recovering the unknown pure component factors C and S from the measured D is a so-called inverse problem, which is closely related to the blind source separation problem [31] or the cocktail party problem [32]. Only matrices C and S with nonnegative entries can make physical sense. Thus $D = CS^T$ poses a nonnegative matrix factorization problem for the given matrix D. See the monographs [9, 10] for a detailed introduction to MCR for applications in chemistry.

The nonnegative matrix factorization problem is not easy to solve. A main hurdle is the non-uniqueness of the solution of such factorization problems. In order to explain this non-uniqueness let $D = CS^T$ be a nonnegative matrix factorization (NMF) of D which means that $C, S \geq 0$. Typically many invertible matrices $R \in \mathbb{R}^{s\times s}$ exist so that

$$D = CS^T = (CR^{-1})(RS^T) = \widetilde{C}\widetilde{S}^T \tag{2}$$

is also a nonnegative factorization of D with the factors $\widetilde{C}$ and $\widetilde{S}$. Trivial examples of such R are those that can be formed as a product of a permutation matrix and a diagonal matrix with positive diagonal elements. These matrices are called generalized permutation matrices and effect simultaneous permutations of the columns of C and S together with mutual up- and down-scaling of the columns. These operations do not yield any additional information on the pure components, but correspond to an interchange of the components together with a scaling. More importantly, often other feasible matrices R exist that are different from generalized permutation matrices. From the viewpoint of the chemical MCR problem this solution ambiguity is very disturbing as only very few (if not only one) factorizations exist which have a chemical meaning. In the research field of chemometrics this problem of non-uniqueness is known under the keyword of *rotational ambiguity*, see [11, 10, 19, 26, 29] among many others.

3.2. Reconstruction by singular value decomposition

Many methods for estimating the pure components are based on a singular value decomposition (SVD) of D, see [33]. The SVD $D = U\Sigma V^T$ is a three-factor decomposition of D where $U \in \mathbb{R}^{k\times k}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{k\times n}$ is a (rectangular) diagonal matrix. The diagonal elements of Σ are the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$. The columns of U are called the left singular vectors of D and the columns of V are called the right singular vectors of D.

If D has the rank s, then only its first s singular values are positive and the remaining singular values equal zero. Then D can be represented only by the first s left and the first s right singular vectors and the associated singular values; for small s this corresponds to an effective data compression. In the case of noisy experimental data the rank-s matrix D is perturbed by noise $\widetilde{D} = D + E$. Typically, $\widetilde{D}$ has the maximal rank $\operatorname{rank}(\widetilde{D}) = \min(k, n)$. If the noise level is small, then its first s largest singular values dominate the remaining singular values which are close to zero. The SVD is the starting point for the construction of noise-filtered low rank approximations of D. Only the first s singular vectors carry important structure information and the remaining singular vectors with their associated close-to-zero singular values can be ignored. This amounts to a low-rank approximation of $\widetilde{D}$ by D. For data sets 1 and 2 the two figures 3 and 4 show the largest twenty singular values and the singular vectors belonging to the dominant singular values.

For experimental spectral data the size distribution of the singular values helps to determine the number of dominant, structurally important singular values s. This number is the barrier to the remaining close-to-zero singular values that have their origin in noise. Usually, s is much smaller than k and n. Then the so-called *truncated SVD* is formed only

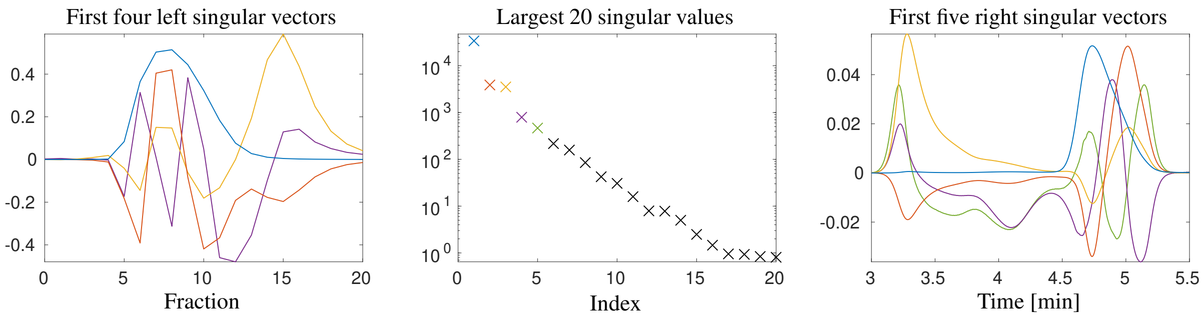


Figure 4: The dominant singular values and associated singular vectors by an SVD of the matrix underlying data set 2. Three dominant singular values can be assumed, but there is no clear break-point between the dominant singular values and those close-to-zero singular values which are to be associated with noise.

by the s dominating singular values and the associated singular vectors. By using the same variable names also for the truncated matrices, namely $U \in \mathbb{R}^{k\times s}$, $V \in \mathbb{R}^{n\times s}$ and $\Sigma \in \mathbb{R}^{s\times s}$, the factorization approach (2) finds its pendant in

$$D = U\Sigma V^T = \underbrace{U\Sigma T^{-1}}_{C}\,\underbrace{TV^T}_{S^T} = CS^T. \tag{3}$$

Therein T is an invertible s-by-s matrix; for details see [11, 10, 34]. The main advantage of this approach is the reduction of the degrees of freedom. A direct computation of C and S includes $(n + k)s$ variables. In contrast to this, the representation of (3) reduces the degrees of freedom to the s^2 matrix elements of T. Equation (3) is the pivotal point for the optimization-based solution of MCR problems by using penalty and regularization functions in Sec. 3.3 and for the general analysis of the set of all nonnegative factorizations of D in Sec. 3.4.

3.3. Optimization-based solution of the MCR problem

The computation of meaningful, chemically interpretable pure component factorizations is by no means trivial [18]. Sometimes additional information on the expected structure of the pure component factors is available which can help to steer the factorization process in a way that the pure component profiles are meaningful and interpretable. Such a control of the factorization process can be achieved by a weighted sum of penalty and regularization terms [35, 12, 34, 16]. The idea is to minimize the reconstruction error $D - CS^T$ under the constraints that C and S are nonnegative matrices plus further constraints. Technically, this is implemented by considering the optimization problem

$$f(T) \to \min \quad \text{with} \quad f(T) = \|D - CS^T\|_F^2 + \alpha_1\|\min(0,C)\|_F^2 + \alpha_2\|\min(0,S)\|_F^2 + \sum_{i=1}^{p}\beta_i g_i(C,S)^2. \tag{4}$$

Therein, $\|\cdot\|_F^2$ denotes the Frobenius norm, which is the sum of squares of all matrix elements of its matrix argument. The coefficients $\alpha_1, \alpha_2 > 0$ are weight factors that determine how strongly the nonnegativity of C and S is enforced. Further regularization terms g_i, $i = 1, \dots, p$, with weights $\beta_i \geq 0$ and $\beta_i \ll \alpha_1, \alpha_2$ can be added.

Due to the representation of C and S by (3) the residual term $\|D - CS^T\|_F^2$ can be replaced by $\|I_{s\times s} - T^{-1}T\|_F^2$ without changing the result of the optimization since $\|D - CS^T\|_F^2 = \|I_{s\times s} - T^{-1}T\|_F^2 + \sum_{i=s+1}^{\min(k,n)} \sigma_i^2$ and the latter term is a fixed value not influencing the minimum point. Therein $I_{s\times s}$ denotes the s-by-s identity matrix. In the case of noisy data a strict penalization of negative entries of C or S can be too rigorous. Instead, small positive control parameters $\varepsilon_C, \varepsilon_S$ can be used that define lower boundaries on acceptable negative entries of C and S in the form of relative size limits

$$\frac{\min C(:,i)}{\max C(:,i)} \geq -\varepsilon_C, \quad \frac{\min S(:,i)}{\max S(:,i)} \geq -\varepsilon_S, \qquad i = 1\dots, s. \tag{5}$$

The numerical minimization of the objective function f by (4) can be achieved by optimizers as the FORTRAN routine NL2SOL by ACM or the MatLab routine lsqnonlin. Sometimes a pre-optimization by a genetic algorithm is worthwhile. If the data includes not only small deviations from the bilinear model (1), then it can be useful to use a higher number of singular vectors $z > s$ in order to reconstruct the s pure component profiles [36, 34]. Such an approach is used if significant information on certain profiles cannot be extracted only from the first s singular vectors. Later we use this technique for data set 2.

The final success of such a Pareto optimization approach depends first and foremost on a proper selection of the regularization functions and their weighting. An imbalanced weighting can prevent convergence to meaningful solutions. For strongly rectangular data sets with $k \ll n$ or $n \ll k$, as this is the case for data set 2, the weight factors should be adapted properly. The unimodality constraint is of special importance for the analysis of chromatographic profiles in this work. The choice of the weighting constants is to be discussed later. Finally, we apply later the *L-curve* technique [37, 38] that serves to visualize the trade-off between the objective function and constraint functions in dependence on the choice of the weight factors.

For the given biopharmaceutical data we expect unimodal profiles in various instances. For example, the elution profile of a single component typically has a *one-maximum* behavior, which means that it is continuously decreasing on both sides of the maximum. For noisy and perturbed experimental data we are interested in profiles that are more or less unimodal. Thus we need a method that allows us to judge to which extent a certain profile is close to a unimodal profile. Algorithm 1 suggests a solution in the form of a simple MatLab-code that evaluates profiles with respect to their closeness to a unimodal shape. Figure 5 shows a pure component estimation for the data set 1 under the constraint that the factor C consists of nearly unimodal profiles. Obviously the blue and the ochre elution profiles are not perfectly unimodal. However, the deviations from unimodality appear to be acceptable for experimental (non-model) data.

Algorithm 1: MatLab-code for the evaluation how close a certain profile $c = C(:, i)$ is to a unimodal profile. The control parameter $\omega \geq 0$ controls the acceptable deviation from unimodality. The residual vector $r \in \mathbb{R}^{k-1}$ has negative entries when $C(:, i)$ breaks the ω-weakened unimodality constraint

```
function unival = unimodeval(c,omega)
r = zeros(size(c));
[cm,i0] = max(c);
for j=i0:-1:2
    r(j-1) = min(0, cm-c(j-1)+omega);
    if ((cm > c(j-1)) || (cm-c(j-1)+omega < 0))
        cm = c(j-1);
    end
end
cm = c(i0);
for j=i0+1:length(c)
    r(j-1) = min(0,cm-c(j)+omega);
    if ((cm > c(j)) || (cm-c(j)+omega < 0))
        cm = c(j);
    end
end
unival = norm(r,2);
```

Other typical regularizations are those that support smooth profiles, those with a small integral (by computing the Euclidean norm of the profile) or regularizations that foster profiles with a high selectivity. A further very effective regularization supports the concentration profiles to be close to the concentration profiles resulting from a given kinetic model [39, 40, 10, 41, 42].

Next we estimate the pure components for data set 1 assuming a number of $s = 3$ components. Unimodality is a constraint for the elution profiles. The control parameters and weight factors are $\varepsilon_C = 7 \cdot 10^{-3}$, $\varepsilon_S = 10^{-4}$, $\alpha_1 = \alpha_2 = 1$ and $\beta_{\text{unimodal}, C} = 0.1$. The control parameter for the unimodality constraint is $\omega = 10^{-4}$. The resulting function values of the unimodality penalty functions are (without weight factors)

$$g_{\text{unimodal, ribonuclease A}} = 8.40 \cdot 10^{-5}, \quad g_{\text{unimodal, cytochrome c}} = 3.15 \cdot 10^{-6}, \quad g_{\text{unimodal, lysozyme}} = 2.80 \cdot 10^{-7}.$$

All other penalty functions have function values that are smaller than $3 \cdot 10^{-11}$. The results are plotted in Fig.5.

3.4. Low-dimensional representation of profiles by the AFS

As already mentioned, the spectral data matrix D has nearly always many nonnegative matrix factorizations, but only one, namely the chemically correct factorization, is of interest for the chemist. There is no golden way from the data to a single and chemically most meaningful solution. A first step towards such a solution is to determine all possible NMFs of the given observed data matrix. This provides an overview on the solution space from which the chemist can typically exclude chemically non-interpretable profiles. In ideal cases this process can result in the desired chemically correct factorization. The benefit of this approach is its reliability and verifiability - in contrast to

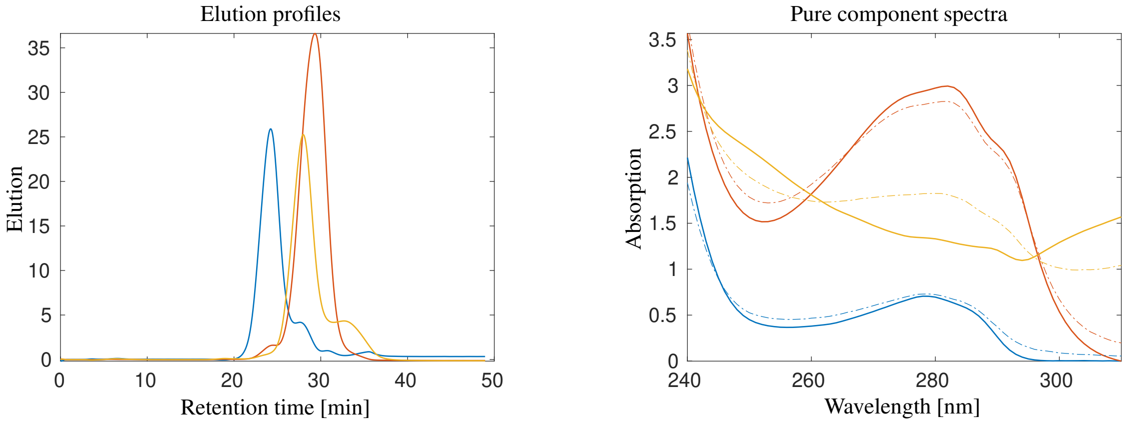


Figure 5: A pure component calculation for data set 1 under the constraint that the columns of C should be close to unimodal profiles. The elution profiles (left) and implicitly the spectra (right) are scaled in a way to achieve a minimal distance to the given profiles (drawn as dash-dotted lines) of the proteins ribonuclease A (blue), lysozyme (red) and cytochrome c (ochre). The known profiles are not used for the computation of the factorization.

this, MCR methods, which result in a single and often method-dependent solution, tend to bias the solution selection by additional uncertain assumptions. As we are now interested in the set of *all* nonnegative factorizations of D, the question is how to make accessible the sets of all matrix factors C and S and how to represent these sets. The answer is simple: We are only interested in the sets of all possible columns of C and S that can be extended to a nonnegative matrix factorization of D. These sets of profiles can be represented in the form of bands of feasible profiles or in a low-dimensional way by plotting their expansion coefficients with respect to the bases of left and right singular vectors. We take the latter option below.

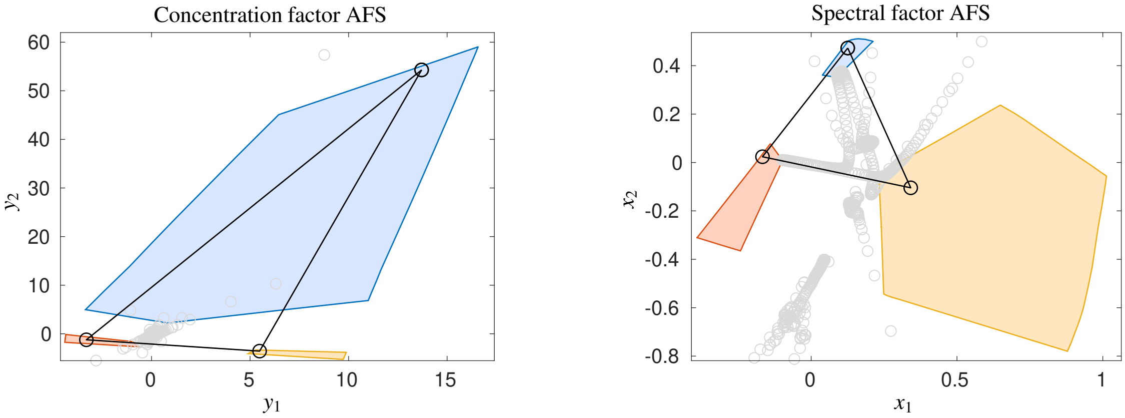


Figure 6: The AFS-sets (colored red, blue and ochre) for the concentration factor and the spectral factor for data set 1. Each point in the AFS defines a feasible profile. The vertices of the two triangles (by black solid lines) mark the chemically correct profiles. The small gray circles represent the columns and rows of D in the AFS. For theoretical reasons, see for example [29], and for noise-free data, all these gray circles must be enclosed by the two triangles. For the given experimental data set very small negative matrix entries are acceptable; we used the control parameters $\varepsilon_C = \varepsilon_S = 7 \cdot 10^{-3}$ by Eq. (5).

3.4.1. The area of feasible solutions (AFS)

Equation (3) has a key function for the low-dimensional representation of the feasible profiles since it allows to present the profiles (being high-dimensional vectors) by the few matrix elements of T or its inverse T^{-1}. The matrix elements of T are the expansion coefficients with respect to the basis of right singular vectors and the elements of T^{-1} determine together with the singular values the expansion coefficients with respect to the basis of left singular vectors. This allows us to define the AFS. For an s-component system the spectral factor AFS is a subset of the $(s-1)$-dimensional space and reads

$$\mathcal{M}_S = \left\{ x \in \mathbb{R}^{s-1} : \quad \text{exists } T \in \mathbb{R}^{s \times s} \text{ with } T(1,:) = (1, x^T),\ \text{rank}(T) = s \text{ and } C, S \geq 0 \right\} \tag{6}$$

with C and S as in (3). The details underlying this definition, especially the leading 1's in each row of T, are rather technical and skipped here. We refer for detailed explanations to [24, 26, 27] and the survey contributions [43, 28, 29].

The associated AFS $\mathcal{M}_C$ for the concentration factor is defined in an analogous way [44, 29]. Further, $\mathcal{M}_C$ equals the spectral AFS of the transposed matrix D^T since in $D^T = SC^T$ the matrices C and S have changed their places compared to $D = CS^T$. In recent years various methods have been developed for the geometric construction and the numerical approximation of the AFS, see e.g. [45]. Of key importance for these developments are the Perron-Frobenius spectral theory of nonnegative matrices [46], the crucial nonnegativity constraints [24, 47], the representation of the single rows and columns of D in the AFS and the associated simplex constructions [24, 44].

For the data sets 1 and 2 the AFS is computed by the polygon inflation algorithm [27]. Each AFS consists of three separated subsets. For this experimental data small deviations from strict nonnegativity should be allowed. The control parameters are $\varepsilon_C = \varepsilon_S = 7 \cdot 10^{-3}$ according to (5). Fig. 6 shows the concentration factor and the spectral factor AFS for data set 1 together with the true profiles by Fig. 5, namely those elution profiles with the smallest deviation from unimodal profiles and those spectra that are closest in the mean squares sense to the given spectra. These six profiles are represented by the six vertices of the two triangles.

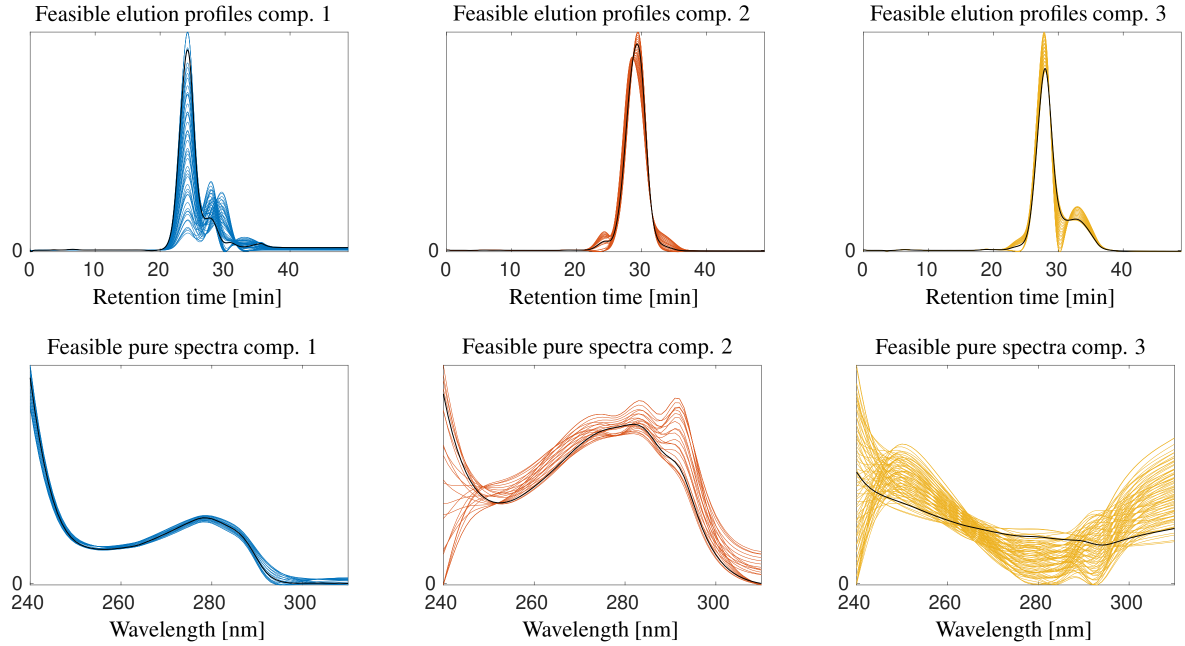


Figure 7: Feasible elution profiles (top) and pure component spectra (bottom) for data set 1. These profiles correspond to the AFS sets as shown in Fig. 6. The finitely many profiles in each plot correspond to the same number of the grid points of uniform grids covering the AFS subsets. No further requirements were placed on these profiles than nonnegativity. Many of the elution profiles are not unimodal and are thereby not meaningful. Fig. 8 shows the remaining profiles under unimodality constraints. The true profiles from Fig. 5 are plotted as black lines.

The mathematical relation between points in the AFS, namely vectors of expansion coefficients, and the associated profiles is as follows: A point $x^{(i)} = (x_1^{(i)}, x_2^{(i)}) \in \mathcal{M}_S$ or $y^{(i)} = (y_1^{(i)}, y_2^{(i)}) \in \mathcal{M}_C$ represents the associated (feasible) profile

$$s^{(i)} = (1,\, x^{(i)})V^T \quad \text{or} \quad c^{(i)} = U\Sigma(1,\, y^{(i)})^T.$$

If we cover an isolated subset of the AFS with a more or less uniform grid and plot for each node of the grid the associated profile, then we get a good approximation of the band of feasible profiles. Figure 7 shows these bands of profiles for the two AFS sets as shown in Fig. 6.

3.4.2. The AFS under unimodality constraints

Assuming elution profiles as unimodal is often well-justified and has a strong impact on the set of feasible solutions. This is reflected in smaller bands of feasible solutions and in a smaller AFS in comparison to the non-reduced original AFS. The idea to combine unimodality assumptions with the AFS is relatively new and discussed in [48, 49].

Figure 8 shows the unimodality restricted AFS and the associated profiles for data set 1. For these computations the control parameters on nonnegativity according to Eq. (5) are $\varepsilon_C = \varepsilon_S = 7 \cdot 10^{-3}$. The control parameter $\omega = 0.025$ limits small deviations from strict unimodality, cf. Algorithm 1. The weight factors are $\alpha_1 = \alpha_2 = 1$ and $\beta_{\text{unimodal, C}} = 0.1$. The unimodality restrictions on the concentration factor AFS imply, by so-called duality, further restrictions on the spectral AFS; we skip the plots of this duality-based restricted spectral AFS here.

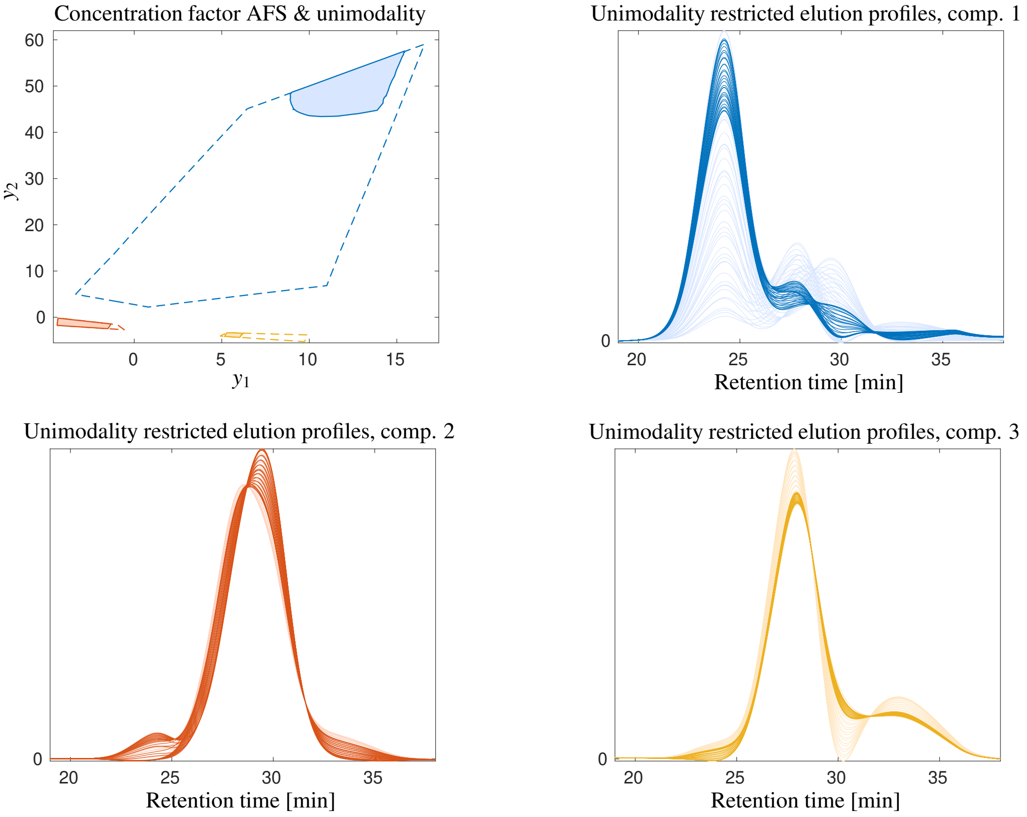


Figure 8: Impact of the unimodality constraint on the AFS and the bands of feasible elution profiles for data set 1. Top left: Reduced AFS. The original AFS is marked by dashed lines. Other plots: The associated bands of unimodal elution profiles. The original profiles are marked in transparent colors in order to point out the effectiveness of the unimodality constraint.

3.4.3. Problems with bi-directional unimodal profiles

The AFS computation for data set 2 yields AFS sets of special structure. The concentration factor AFS and the spectral factor AFS consist of very large and narrow subsets together with others that are much smaller and close to the origin. All this makes it difficult to construct meaningful solutions in the geometric way as explained above. We see a possible explanation in the structure of D as shown in the left subplot of Fig. 2 where two main nonzero signal regions can be identified that are weakly coupled. Weakly coupled sub-spectra suggest to split the data set to two independent data sets and then to subject them to a separate analysis. On the one hand, this would ignore the weak coupling. On the other hand, a joint analysis of the full data set runs into the problem that the matrix $D^T D$ is close to a reducible matrix for which the AFS cannot work. In such borderline cases, it is known that the AFS can consist of long narrow subsets. A possible explanation is that unimodal profiles are to be constructed from linear combinations of singular vectors. Only the first singular vector can be unimodal, whereas the other singular vectors are typically due to their mutual orthogonality non-unimodal. The finding of unimodal linear combinations is then difficult - a circumstance which can intuitively be understood while working with the FACPACK module Duality/Complementarity & AFS on this data set; see Appendix A on the FACPACK software.

4. Results & Discussion of the chemometric analysis

Next further results on the pure component estimation for the two data sets are presented and discussed. First, we consider the pure component estimation of the problematic data set 2 under unimodality constraints for each of the two factors, see also Sec. 3.4.3. Second, we discuss the question of how to select a proper number of components s. Various estimations are computed and analyzed.

The results in this section demonstrate the potential of AFS techniques for analyzing ambiguous MCR outputs. We do not claim that the AFS can help for any liquid chromatography separation. However, the usefulness of AFS techniques is likely to increase with an increasing complexity and mutual overlap of the given chromatographic data.

4.1. Pure component calculation for data set 2

As already mentioned in the first paragraph of Sec. 3, the naming of the second factor in $D = CS^T$ suggests an interpretation as a spectral factor. However, data set 2 does not include a frequency axis as all absorption measurements are taken at the fixed wavelength 280 nm. As we focus on the bilinear data structure we do not see a strong necessity for introducing a new variable name for the second factor. Unimodal profiles are expected for each factor. Within the factorization process it turned out that two observable peaks in the analytical dimension have nearly the same preparative elution profile. Hence it can be merged in a joint analytical elution profile, which is then a bi-modal profile. Thus unimodality is applied only for all but one analytical and for all preparative elution profiles. For the reconstruction of $s = 3$ components the information from $z = 4$ singular vectors is used, see [34]. The control parameters and weight factors are $\varepsilon_C = \varepsilon_S = 10^{-4}$, $\alpha_1 = 100$, $\alpha_2 = 1$, $\beta_{\text{unimodal, C}} = 10$, $\beta_{\text{unimodal, S}} = 50$ and $\omega = 10^{-4}$. The considerably different sizes of the weight factors is justified by the asymmetric resolution for the chromatograms, namely $k = 21$ and $n = 3751$. Additionally, the function values of the function penalizing deviations from unimodality are usually much smaller than those on deviations from nonnegativity. In order to balance these constraints, the weight factor of the unimodality constraint function is taken larger than that of the nonnegativity constraint function. Whether or not the weight factor choice is suitable can be checked by evaluation of the weighted residuals. For the given problem well-balanced values read as follows

$$\|D - CS^T\|_F^2 - \sum_{i=4}^{21} \sigma_i^2 = 1.3 \cdot 10^{-29}, \qquad \alpha_1^2 g_{\text{nonneg}}(C) = 2.1 \cdot 10^{-2}, \qquad \alpha_2^2 g_{\text{nonneg}}(S) = 5.0 \cdot 10^{-2},$$

$$\beta^2_{\text{unimodal, C}} g_{\text{unimodal}}(C) = 7.4 \cdot 10^{-2}, \qquad \beta^2_{\text{unimodal, S}} g_{\text{unimodal}}(S) = 1.4 \cdot 10^{-3}.$$

The results are presented in Fig. 9 for the case of selecting $s = 3$ main components. From the perspective of a practitioner, the species drawn in blue and red in Fig. 9 correspond to the aggregates and the product, respectively. The elution profiles are largely consistent with the results from manual data analysis shown in Fig. 5B in [30]. The species drawn in ochre was not detected during manual data analysis in [30]. It seems to correspond to a lumped component of non-ideal elution behavior of the main peak and of product fragments which tend to be obscured by the main peak. Upon careful inspection of the analytical chromatograms in the right subplot in Fig. 2, a shoulder is distinguishable above 5 min retention time. While a manual analysis would be challenging, MCR allows to systematically extract the shoulder as additional species thus improving reproducibility.

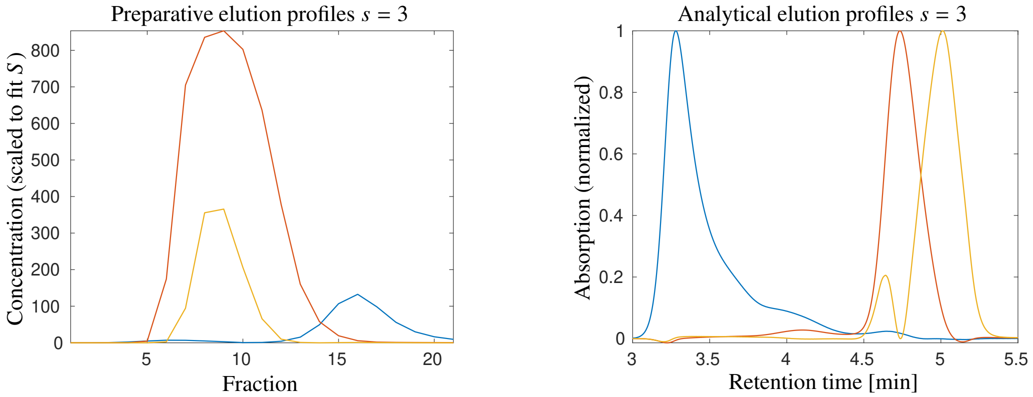


Figure 9: Pure component estimation for data set 2 for the case of $s = 3$ assumed species. Unimodal profiles are used for each factor except for one analytical elution profile drawn in ochre for which the singular value decomposition does not permit a signal separation. This compound profile includes the two peaks centered at 4.64 min and 5.01 min. For the reconstruction $z = 4$ singular vectors are used in order to improve the results.

4.2. Verification of the weight factor ratios by an L-curve for data set 2

A suitable choice of the weight factors is important for the final success of the Pareto optimization of the objective function (4). Next various combinations of weight factors are tested for data set 2. Then the resulting pure component profiles are used in order to evaluate the penalty functions including their weight factors in separate form. For these computations we assume $s = 3$ components.

Next we compare the nonnegativity constraint and the unimodality constraint. The weight factors are as follows

$$\alpha_1 = 100, \quad \alpha_2 = 1, \quad \alpha_3 = 1, \quad \beta_{\text{unimodal, C}} = \gamma, \quad \beta_{\text{unimodal, S}} = 5\gamma. \tag{7}$$

The L-curve is plotted in Fig. 10 for several values $\gamma \in [0.01, 500]$, see [37, 38] for the L-curve technique.

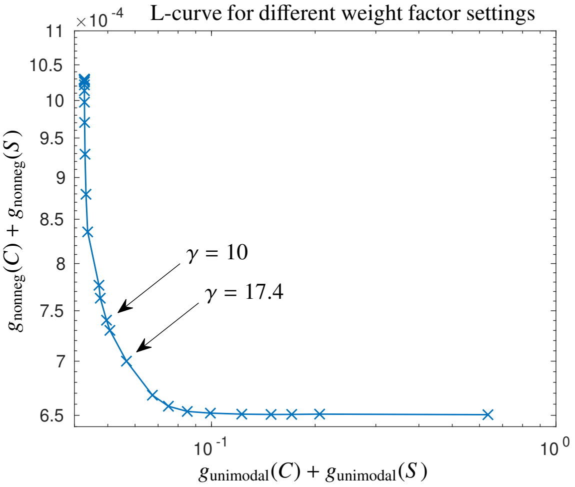


Figure 10: L-curve for various weight factor settings for data set 2 and $s = 3$ components. The choice $\gamma = 10$ in (7) results in the estimation shown in Fig. 9. Values of γ that are closer to the kink (point of highest curvature) of the L-curve as $\gamma = 17.4$ result in a nearly unchanged factorization. The total deviation is $\|C - \widetilde{C}\|_F^2/\|C\|_F^2 + \|S - \widetilde{S}\|_F^2/\|S\|_F^2 = 10^{-5}$ if (C, S) correspond to $\gamma = 10$ and $(\widetilde{C}, \widetilde{S})$ correspond to $\gamma = 17.4$.

4.3. Detection of the number of components

The number s of singular vectors that are used for the construction of the pure component factors is sometimes difficult to determine. Indicators on the size of s can be a prior knowledge on the number of chemical species in the system or the size distribution of the singular values of the matrix D. In ideal cases the size distribution shows a clear break between a dominant set of s singular values that are characteristically larger than the remaining set of singular values. These latter singular values are sometimes close to zero and their origin can be seen in noise. Typical situations are illustrated by Fig. 5 in [29], Fig. 2 in [50] or Fig. 5 in [41].

The decision on a proper choice of s is problematic for the given protein chromatographic data sets, cf. the plots of the singular values in Figs. 3 and 4. For data set 1 either $s = 3$ or $s = 4$ are reasonable values. For data set 2 the choice $s = 3$ appears to be possible for a first analysis, but this ignores many singular values that are relatively large. The observed high number of relevant singular values was attributed to multiple sources. First of all, protein systems generally contain a large number of subspecies, i.e. product- and contaminant-isoforms. Furthermore, analytical chromatography may display deviations from ideal bilinear behavior [8]. As we try to reconstruct only the main species and neglect analytical imperfections, the SVD helps to denoise the data.

4.3.1. Pure component estimation for data set 1 with $s = 4$ species

Whenever there are doubts about the appropriate choice of the number of components s, it is advisable to experiment with an increased or sometimes decreased s. Therefore we try a second estimation for data set 1 with $s = 4$ components. Once again unimodality constraints are applied to the elution profiles. The results are presented in Fig. 11 together with a comparison to the results for $s = 3$. The estimation for $s = 4$ species is not interpretable since only the component ribonuclease A (blue) appears to be correct and the other spectral profiles cannot be associated to the other protein species. Especially the red spectrum shows pronounced deviations from the typical profile with a minimum near 280 nm instead of a maximum.

4.3.2. Pure component factorization for data set 2 and different numbers of species s

The SVD of the matrix underlying data set 2 does not clearly indicate the number of species s. We tried various combinations of the numbers s and z. Unimodality constraints are used for all but one profile in the analytical elution direction and for all profiles in the preparative elution direction. The weight factors for all these experiments are listed in Sec. 4.1. Decompositions are computed for $s \in \{2, \ldots, 7\}$. The resulting minimal values of the objective function (4) are numerically evaluated. These values are plotted in Fig. 12 together with the final values of the unimodality constraint function and the relative reconstruction errors. Only $s = 3$ results in an interpretable solution, see Fig. 9. For the factorization with $s = 2$ a number of $z = 4$ singular vectors has been used. For all other factorizations z equals $s + 1$.

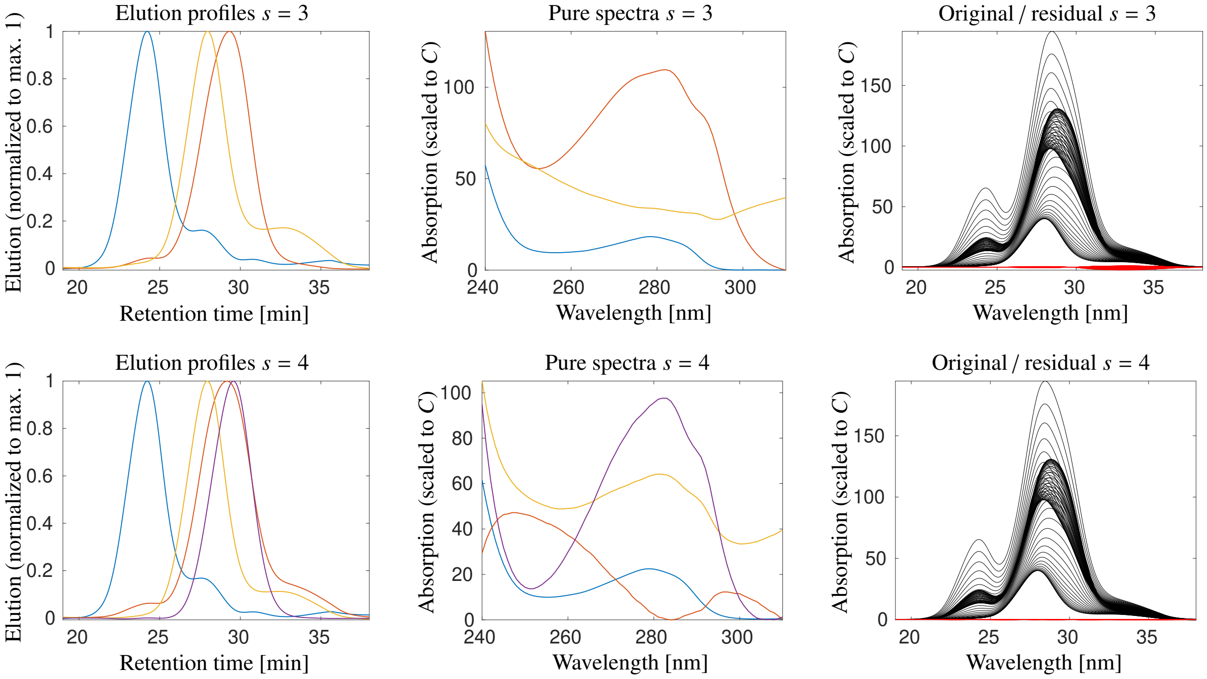


Figure 11: Pure component estimations and the residuals for data set 1 with respect to $s = 3$ (top row and see also Fig. 5) and $s = 4$ (lower row). For the ($s = 4$)-component estimation only the pure component of ribonuclease A (blue) is correct. In contrast to this, all profiles of the estimation with $s = 3$ components can be associated with biopharmaceutical components. In the right plots the residuals $D - CS^T$ are drawn in red together with the original spectral data in black. Not surprisingly the residual is smaller for the larger $s = 4$.

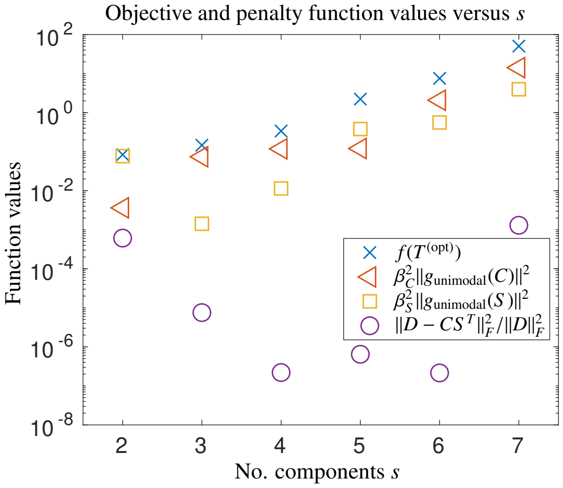


Figure 12: Function values of the objective functions and of the penalty functions for pure component estimations of data set 2 versus various settings of the number of components s. For each $s \in \{2, \ldots, 7\}$ the objective function f by (4) is optimized and the final values for the optimal T are evaluated and plotted by blue crosses. Further, the final values of the penalty function on unimodality for the preparative elution profiles are marked by red triangles. The analogous values for the analytical elution profiles are plotted by ochre squares. The weight factors are $\beta_C = 10$ and $\beta_S = 50$. Additionally, the relative reconstruction error $\|D - CS^T\|_F^2/\|D\|_F^2$ is shown by purple circles. However, the latter term is not part of the objective function in this form. Therein $C = U\Sigma(T^{(\text{opt})})^+$ and $S^T = T^{(\text{opt})}V^T$ are the factors for the optimal transformation $T^{(\text{opt})}$. Biochemically interpretable are only the results for the choice $s = 3$. See Fig. 9 for the resulting unimodal profiles for $s = 3$ and $s = 4$.

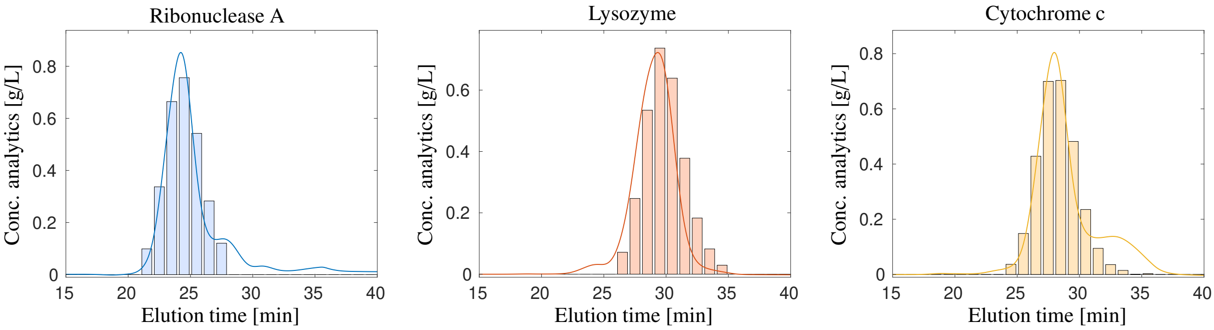


Figure 13: Comparison of the computed results (lines) and the off-line analytic (bars, measured concentrations) for data set 1. The profiles are scaled in terms of a least-squares fit compared to the offline analytic. The deviations appear to be acceptable with respect to the different measurements. Off-line analytics results in profiles with a slight tailing. The peaks are asymmetric for higher retention times.

4.4. Verification of the results by quantitative analysis

The methodology introduced and applied in this paper results in qualitative (non-scaled) profiles. The knowledge of an underlying kinetic model, additional information of certain concentration profiles or of elution profiles (for example from separate measurements) can help to determine the correct scaling of all profiles forming the columns of C and S. Although a quantitative analysis is not a primary focus of the suggested techniques, we carry out a quantitative analysis to check the accuracy of the results for data set 1.

Fig. 13 compares the computed profiles with off-line analytical results for data set 1. The profiles are scaled by means of a least-squares fit to offline analytical measurements (interpolation is used due to different time grids). The results are largely consistent. Minor deviations can be traced back to different recording techniques (the values of off-line analytics are averaged values). We note that the measuring sites of the off-line analytics and for recording spectroscopic data D are different. This can result in a minor tailing of the peak profiles and a slight profile smoothing due to dispersion of the chromatographic system.

Next the root mean square error of the prediction (RMSEP), the coefficients of determination (R^2 values) as well as a relative error estimator Δ are evaluated in order to compare the computed and the predicted profiles. These values are given by

$$RMSEP_\ell = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(C(i,\ell)-\widehat{C}(i,\ell)\right)^2}, \qquad R_\ell^2 = 1-\frac{\sum_{i=1}^{n}\left(C(i,\ell)-\widehat{C}(i,\ell)\right)^2}{\sum_{i=1}^{n}\left(C(i,\ell)-\overline{C}_i\right)^2}, \qquad \Delta_\ell = \frac{\sum_{i=1}^{n}\left(C(i,\ell)-\widehat{C}(i,\ell)\right)^2}{\sum_{i=1}^{n}\left(\widehat{C}(i,\ell)\right)^2}$$

for $\ell = 1,\ldots,3$. Therein C are the computed profiles. Further $\widehat{C}$ contains the predicted profiles and the mean values are given by $\overline{C}_\ell = \frac{1}{n}\sum_{i=1}^{n} C(i,\ell)$. The profiles are interpolated for the $n = 49$ predicted grid points. Once again the profiles are scaled in terms of a least squares fit. The values are

$$RMSEP = (0.0373,\ 0.0529,\ 0.0380)\,\text{g/L}, \qquad R^2 = (0.946,\ 0.886,\ 0.942), \qquad \Delta = (0.045,\ 0.092,\ 0.048)$$

in the order ribonuclease A, lysozyme, cytochrome c.

4.5. Comparison with results gained by MCR-ALS

MCR-ALS is a well established MCR-toolbox for the pure component recovery from spectral mixture data, see e.g. [12, 51]. Its idea is to compute a factorization $D = CS^T$ by solving iteratively least squares problems in an alternating way and where nonnegativity constraints are applied in each step. This strategy is comparable to classical numerical routines to compute nonnegative matrix factorizations (NMFs), see [21]. Additional constraints as unimodality, closure or equality can also be applied. We use MCR-ALS for the computation of the factors C and S and assume unimodality of the elution profiles (columns of C). Next the results are compared for the data sets 1 and 2. Figure 14 shows the MCR-ALS profiles by dash-dotted lines in combination with the results gained by Pareto optimization and which are separately plotted in Figs. 5 and Fig. 9. The profiles for each of the data sets nearly coincide. Considering the set of all feasible profiles, we would like to point out that the AFS approach is more general as it informs the user on potential ambiguities of the profiles. It gives access to all solutions that any MCR method can provide. But it is equally important to provide selection strategies to find the correct solution within the bands of feasible solutions. The present results indicate that our optimization-based selection strategy can successfully extract the correct profiles.

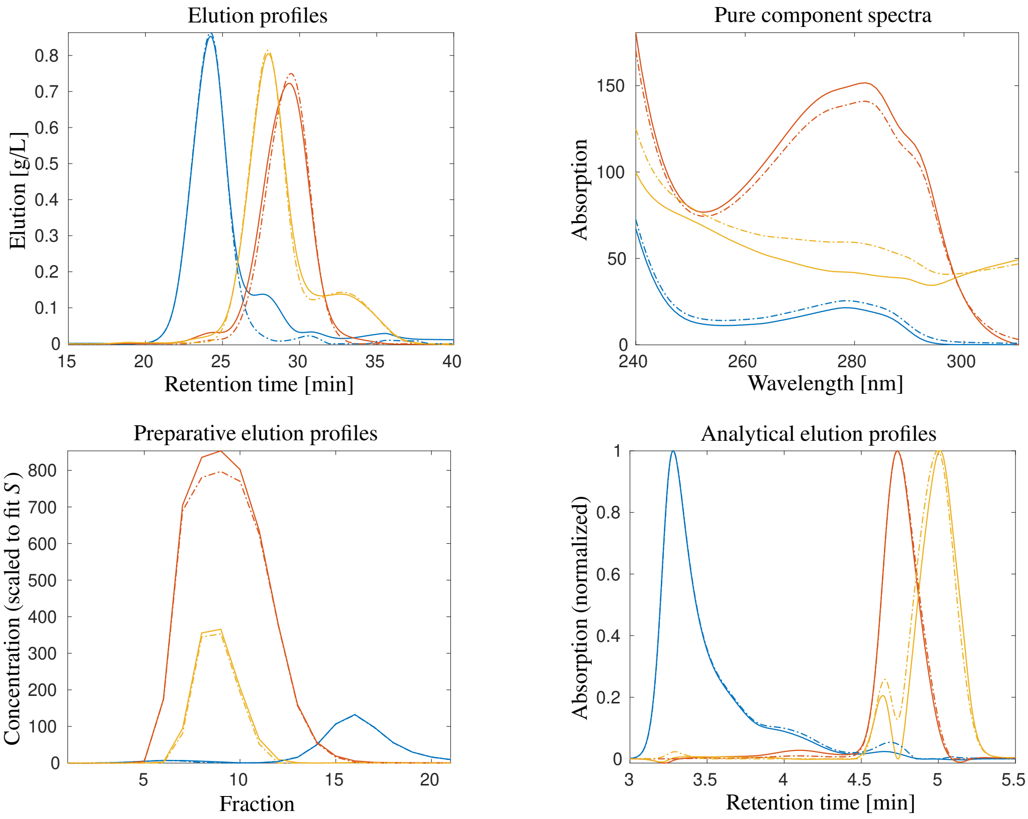


Figure 14: Comparison with MCR-ALS (dash-dotted lines) results for data set 1 (left) and data set 2 (right). The solid lines represent the profiles gained by Pareto optimization. Both computational approaches use unimodality constraints for the concentration profiles.

5. Conclusion

The solution ambiguity underlying MCR methods, although not generally aware, can affect the reliability of chemometric analyses also in the field of biopharmaceutical chromatographic data and protein purification processes. Within the framework of the area of feasible solutions, which contains all possible factorizations, further constraints can help to determine a single, chemically correct solution. The present study demonstrates a successful application of the unimodality constraint in order to steer the pure component factorization in the correct direction. A challenging problem for protein chromatographic data is the correct identification of the number of components that can be reconstructed from the data. Biopharmaceutical protein mixtures containing a larger number of isoforms make it difficult to separate the characteristic structures of the dominating species from other components with smaller contributions.

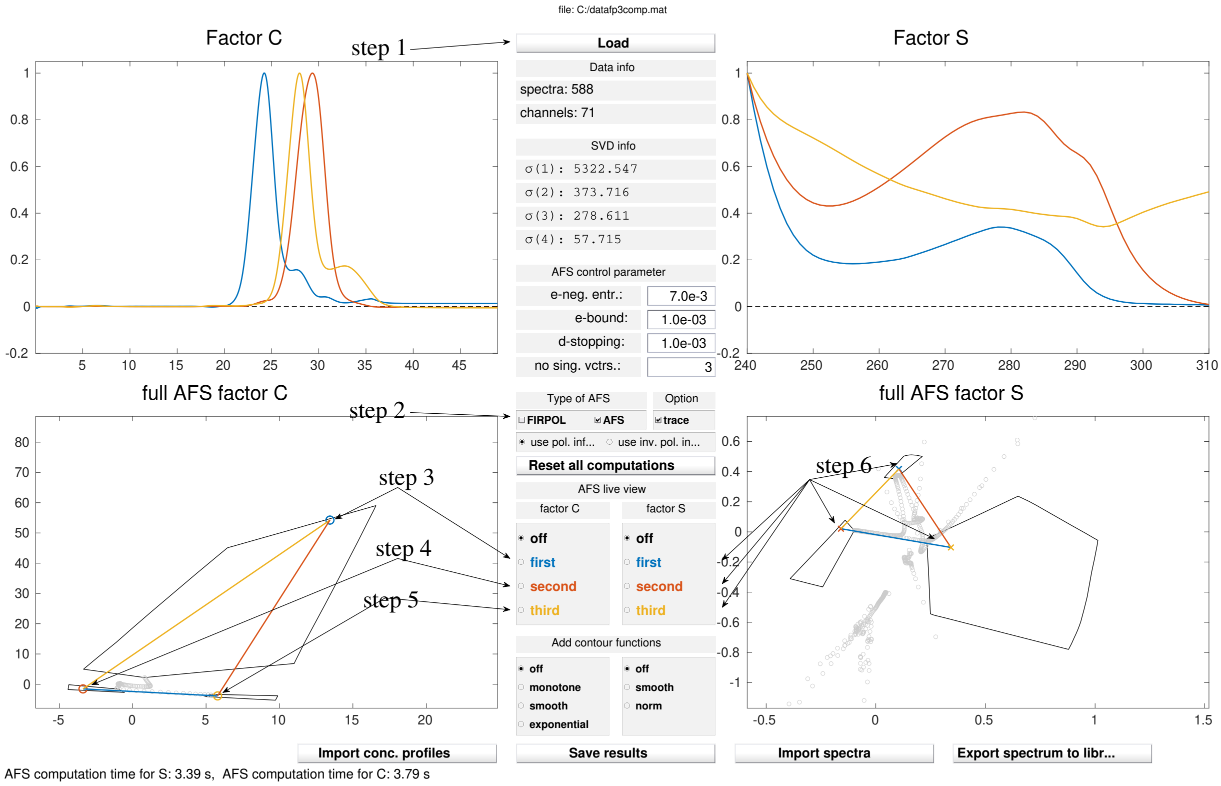
Since unimodality constraints work effectively (at least) for the given chromatographic data, future work may deal with the introduction of additional models on the elution profiles. Similarly, the application of any kinetic modeling, as far as possible, can support a successful pure component factorization. An extension of the MCR software FACPACK by a module that exploits unimodality in one of the factors or simultaneously in both factors is planned.

A. Appendix: AFS computation and profile reconstruction by FACPACK

The software FACPACK [52] is a collection of routines for AFS computations and AFS related data analyses. The terminology and focus is on applications to spectroscopic data on a time×frequency grid. However, any other data with an underlying bilinear structure can also be analyzed. The computational core of the software is written in C and FORTRAN. Using the software is easy since it comes with a MatLab-based graphical user interface. Currently the software is limited to chemical systems with not more than four chemical components ($s \leq 4$) as for systems with a larger number of components a graphical representation of the results in the $(s-1)$-dimensional space is difficult to handle. However, in systems with more than four chemical components a subsystem analysis appears to be possible if a spectral window can be specified in which not more than four components are active.

The module *Duality/Complementarity & AFS (3 components)* of FACPACK gives the user the opportunity to construct pure component decomposition in a step-by-step manner. This can be done by selecting and modifying the vertices of the two triangles representing the pure component factors of a three-component system. Duality relations that express the impact of known columns of C or S^T to the other factor are implemented in an interactive way. This means that if one triangle is modified, then its dual triangle is adapted automatically so that $D = CS^T$ always holds, see the simplex-based geometric construction of pure component decompositions [53, 54, 29]. If for example one vertex of the triangle representing the spectral factor is modified, then this results in a simultaneous visualization of the associated spectral profile and the automatic adaption of the dual vertices of the concentration factor triangle and their associated elution profiles. An activation of the live-view mode of this module makes it possible to see all these modifications as a simultaneous response on the movement of the certain point of the AFS by mouse motions. Presumably correct profiles can be locked by pressing the left mouse button and further profiles can be edited in the same way. Fig. 15 shows a screen shot of the duality module for three-component systems in its application to data set 1.

shot of the


file: C:/datafp3comp.mat

Data info
spectra: 588
channels: 71

SVD info
σ(1): 5322.547
σ(2): 373.716
σ(3): 278.611
σ(4): 57.715

AFS control parameter
e-neg. entr.: 7.0e-3
e-bound: 1.0e-03
d-stopping: 1.0e-03
no sing. vctrs.: 3

Type of AFS: □ FIRPOL ☑ AFS
Option: ☑ trace
● use pol. inf... ○ use inv. pol. in...

AFS live view
factor C: ● off ○ first ○ second ○ third
factor S: ● off ○ first ○ second ○ third

Add contour functions
● off ○ monotone ○ smooth ○ exponential
● off ○ smooth ○ norm

AFS computation time for S: 3.39 s, AFS computation time for C: 3.79 s

Figure 15: Screenshot of the FACPACK module Duality/Complementarity & AFS for three-component systems. Its main operation steps are as follows: Step 1 to load the data. Step 2 to compute the AFS for the concentration factor and the spectral factor. Step 3 to select by choice of the user a first profile, to move the mouse pointer through the associated AFS and to lock a presumably correct profile. Further steps 4 and 5 are to select and to modify the second and third profile. Modification steps can be repeated in order to refine the profile selection. In an optional step 6 the user can apply similar changes to the factor S.

References

[1] G. Carta and A. Jungbauer. *Protein chromatography: process development and scale-up*. John Wiley & Sons, 2010.

[2] H. Schmidt-Traub, M. Schulte, and A. Seidel-Morgenstern. *Preparative chromatography*. Wiley, Weinheim, 2012.

[3] G. Jagschies, E. Lindskog, K. Lacki, and P. M. Galliher. *Biopharmaceutical Processing: Development, Design, and Implementation of Manufacturing Processes*. Elsevier, 2018.

[4] N. Brestrich, T. Briskot, A. Osberghaus, and J. Hubbuch. A tool for selective inline quantification of co-eluting proteins in chromatography using spectral analysis and partial least squares regression. *Biotechnol. Bioeng.*, 111(7):1365–1373, 2014.

[5] B. A. Patel, A. Gospodarek, M. Larkin, S. A. Kenrick, M. A. Haverick, N. Tugcu, M. A. Brower, and D. D. Richardson. Multi-angle light scattering as a process analytical technology measuring real-time molecular weight for downstream process control. In *Mabs*, volume 10(7), pages 945–950. Taylor & Francis, 2018.

[6] N. Brestrich, M. Ruedt, D. Buechler, and J. Hubbuch. Selective protein quantification for preparative chromatography using variable pathlength UV/Vis spectroscopy and partial least squares regression. *Chem. Eng. Sci.*, 176:157–164, 2018.

[7] L. Rolinger, M. Rüdt, and J. Hubbuch. A critical review of recent trends, and a future perspective of optical spectroscopy as PAT in biopharmaceutical downstream processing. *Anal. Bioanal. Chem.*, 412(9):1–18, 2020.

[8] H. Parastar and R. Tauler. Multivariate curve resolution of hyphenated and multidimensional chromatographic measurements: a new insight to address current chromatographic challenges. *Anal. Chem.*, 86(1):286 – 297, 2014.

[9] E. Malinowski. *Factor analysis in chemistry*. Wiley, New York, 2002.

[10] M. Maeder and Y.M. Neuhold. *Practical data analysis in chemistry*. Elsevier, Amsterdam, 2007.

[11] W.H. Lawton and E.A. Sylvestre. Self modelling curve resolution. *Technometrics*, 13:617–633, 1971.

[12] J. Jaumot, R. Gargallo, A. de Juan, and R. Tauler. A graphical user-friendly interface for MCR-ALS: a new tool for multivariate curve resolution in MATLAB. *Chemom. Intell. Lab. Syst.*, 76(1):101–110, 2005.

[13] M.-H. Kamga, H. Woo Lee, J. Liu, and S. Yoon. Quantification of protein mixture in chromatographic separation using multi-wavelength UV spectra. *Biotechnol. Prog.*, 29(3):664–671, 2013.

[14] F. Dismer, S. Hansen, S. A. Oelmeier, and J. Hubbuch. Accurate retention time determination of co-eluting proteins in analytical chromatography by means of spectral data. *Biotechnol. Bioeng.*, 110(3):683–693, 2013.

[15] Z. Xia, A. Liu, W. Cai, and X. Shao. Band target entropy minimization for retrieving the information of individual components from overlapping chromatographic data. *J. Chromatogr. A*, 1411:110–115, 2015.

[16] M. Rüdt, S. Andris, R. Schiemer, and J. Hubbuch. Factorization of preparative protein chromatograms with hard-constraint multivariate curve resolution and second-derivative pretreatment. *J. Chromatogr. A*, 1585:152–160, 2019.

[17] K.A. Bakeev. *Process analytical technology: Spectroscopic tools and implementation strategies for the chemical and pharmaceutical industries*. John Wiley & Sons, 2010.

[18] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5:111–126, 1994.

[19] H. Abdollahi and R. Tauler. Uniqueness and rotation ambiguities in multivariate curve resolution methods. *Chemom. Intell. Lab. Syst.*, 108(2):100–111, 2011.

[20] D.D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[21] H. Kim and H. Park. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. *SIAM J. Matrix Anal. Appl.*, 30:713–730, 2008.

[22] A. de Juan, Y. Vander Heyden, R. Tauler, and D. L . Massart. Assessment of new constraints applied to the alternating least squares method. *Anal. Chim. Acta*, 346(3):307–318, 1997.

[23] S. Vali Zade, K. Neymeyr, H. Abdollahi, and M. Sawall. On the signal contribution function and the area of feasible solutions under unimodality constraints. Submitted 2020.

[24] O.S. Borgen and B.R. Kowalski. An extension of the multivariate component-resolution method to three components. *Anal. Chim. Acta*, 174:1–26, 1985.

[25] R. Rajkó and K. István. Analytical solution for determining feasible regions of self-modeling curve resolution (SMCR) method based on computational geometry. *J. Chemom.*, 19(8):448–463, 2005.

[26] A. Golshan, H. Abdollahi, and M. Maeder. Resolution of rotational ambiguity for three-component systems. *Anal. Chem.*, 83(3):836–841, 2011.

[27] M. Sawall, C. Kubis, D. Selent, A. Börner, and K. Neymeyr. A fast polygon inflation algorithm to compute the area of feasible solutions for three-component systems. I: Concepts and applications. *J. Chemom.*, 27:106–116, 2013.

[28] A. Golshan, H. Abdollahi, S. Beyramysoltan, M. Maeder, K. Neymeyr, R. Rajkó, M. Sawall, and R. Tauler. A review of recent methods for the determination of ranges of feasible solutions resulting from soft modelling analyses of multivariate data. *Anal. Chim. Acta*, 911:1–13, 2016.

[29] M. Sawall, H. Schröder, D. Meinhardt, and K. Neymeyr. On the ambiguity underlying multivariate curve resolution methods. In S. Brown, R. Tauler, and B. Walczak, editors, *In Comprehensive Chemometrics: Chemcial and Biochemical Data Analysis*, pages 199–231. Elsevier, 2020.

[30] N. Brestrich, M. Rüdt, D. Büchler, and J. Hubbuch. Selective protein quantification for preparative chromatography using variable pathlength UV/Vis spectroscopy and partial least squares regression. *Chem. Eng. Sci.*, 176:157–164, 2018.

[31] A. Cichocki, R. Zdunek, A.H. Phan, and S. Amari. *Nonnegative matrix and tensor factorizations: Applications to exploratory multi-way data analysis and blind source separation*. John Wiley & Sons, 2009.

[32] E. C. Cherry. Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.*, 25(5):975–979, 1953.

[33] G.H. Golub and C.F. Van Loan. *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, 2012.

[34] K. Neymeyr, M. Sawall, and D. Hess. Pure component spectral recovery and constrained matrix factorizations: Concepts and applications. *J. Chemom.*, 24:67–74, 2010.

[35] E. Widjaja, C. Li, W. Chew, and M. Garland. Band target entropy minimization. A robust algorithm for pure component spectral recovery. Application to complex randomized mixtures of six components. *Anal. Chem.*, 75:4499–4507, 2003.

[36] W. Chew, E. Widjaja, and M. Garland. Band-target entropy minimization (BTEM): An advanced method for recovering unknown pure component spectra. Application to the FT-IR spectra of unstable organometallic mixtures. *Organometallics*, 21(9):1982–1990, 2002.

[37] P. C. Hansen. Analysis of Discrete Ill-Posed Problems by Means of the L-Curve. *SIAM Review*, 34(4):561–580, 1992.

[38] H. W. Engl, M. Hanke, and A. Neubauer. *Regularization of inverse problems*, volume 375 of *Math. Appl.* Kluwer Academic Publishers, Dordrecht, 2000.

[39] A. de Juan, M. Maeder, M. Martínez, and R. Tauler. Combining hard and soft-modelling to solve kinetic problems. *Chemom. Intell. Lab. Syst.*, 54:123–141, 2000.

[40] J. Jaumot, P. J. Gemperline, and A. Stang. Non-negativity constraints for elimination of multiple solutions in fitting of multivariate kinetic models to spectroscopic data. *J. Chemom.*, 19(2):97–106, 2005.

[41] M. Sawall, A. Börner, C. Kubis, D. Selent, R. Ludwig, and K. Neymeyr. Model-free multivariate curve resolution combined with model-based kinetics: Algorithm and applications. *J. Chemom.*, 26:538–548, 2012.

[42] H. Schröder, M. Sawall, C. Kubis, D. Selent, D. Hess, R. Franke, A. Börner, and K. Neymeyr. On the ambiguity of the reaction rate constants in multivariate curve resolution for reversible first-order reaction systems. *Anal. Chim. Acta*, 927:21–34, 2016.

[43] M. Sawall, A. Jürß, H. Schröder, and K. Neymeyr. *On the analysis and computation of the area of feasible solutions for two-, three- and four-component systems*, volume 30 of Data Handling in Science and Technology, “Resolving Spectral Mixtures”, Ed. C. Ruckebusch, chapter 5, pages 135–184. Elsevier, Cambridge, 2016.

[44] M. Sawall, A. Jürß, H. Schröder, and K. Neymeyr. Simultaneous construction of dual Borgen plots. I: The case of noise-free data. *J. Chemom.*, 31:e2954, 2017.

[45] A. Jürß, M. Sawall, and K. Neymeyr. On generalized Borgen plots. I: From convex to affine combinations and applications to spectral data. *J. Chemom.*, 29(7):420–433, 2015.

[46] H. Minc. *Nonnegative matrices*. John Wiley & Sons, New York, 1988.

[47] M. Sawall and K. Neymeyr. A fast polygon inflation algorithm to compute the area of feasible solutions for three-component systems. II: Theoretical foundation, inverse polygon inflation, and FAC-PACK implementation. *J. Chemom.*, 28:633–644, 2014.

[48] M. Sawall, N. Rahimdoust, C. Kubis, H. Schröder, D. Selent, D. Hess, H. Abdollahi, R. Franke, Börner A., and K. Neymeyr. Soft constraints for reducing the intrinsic rotational ambiguity of the area of feasible solutions. *Chemom. Intell. Lab. Syst.*, 149, Part A:140–150, 2015.

[49] N. Rahimdoust, M. Sawall, K. Neymeyr, and H. Abdollahi. Investigating the effect of flexible constraints on the accuracy of self-modeling curve resolution methods in the presence of perturbations. *J. Chemom.*, 30(5):252–267, 2016.

[50] H. Schröder, M. Sawall, C. Kubis, A. Jürß, D. Selent, A. Brächer, A. Börner, R. Franke, and K. Neymeyr. Comparative multivariate curve resolution study in the area of feasible solutions. *Chemom. Intell. Lab. Syst.*, 163:55–63, 2017.

[51] J. Jaumot, A. de Juan, and R. Tauler. MCR-ALS GUI 2.0: new features and applications. *Chemom. Intell. Lab. Syst.*, 140:1–12, 2015.

[52] M. Sawall, A. Moog, and K. Neymeyr. FACPACK: A software for the computation of multi-component factorizations and the area of feasible solutions, Revision 1.3. FACPACK homepage: http://www.math.uni-rostock.de/facpack/, 2018.

[53] S. Beyramysoltan, H. Abdollahi, and R. Rajkó. Newer developments on self-modeling curve resolution implementing equality and unimodality constraints. *Anal. Chim. Acta*, 827(0):1–14, 2014.

[54] M. Sawall and K. Neymeyr. On the area of feasible solutions and its reduction by the complementarity theorem. *Anal. Chim. Acta*, 828:17–26, 2014.