Sampling-based computation of the sets of feasible solutions and feasible bands for noisy data

Mathias Sawall[a], Tomass Andersons[a], Chunhong Wei[b], Christoph Kubis[b], Klaus Neymeyr[a,b]

[a] *Universität Rostock, Institut für Mathematik, Ulmenstraße 69, 18057 Rostock, Germany*
[b] *Leibniz-Institut für Katalyse, Albert-Einstein-Straße 29a, 18059 Rostock*

Abstract

Multivariate curve resolution often suffers from solution ambiguity, with many nonnegative factorizations fitting the data equally well. Building on the method of Laursen and Hobolth (2022), we present an efficient sampling algorithm that can handle noisy data even containing negative entries. The algorithm iteratively updates factor columns via affine combinations within a nested loop structure, effectively approximating the sets of feasible solutions, the feasible bands, as well as the dual profiles. We apply the method to two in situ FTIR spectroscopic data sets tracking the decomposition and activation of rhodium carbonyl complexes for the hydroformylation process. Performance and accuracy are compared against established algorithms, demonstrating both robustness and computational efficiency.

Keywords: multivariate curve resolution, sampling, factor ambiguity, feasible bands

1. Introduction

Multivariate curve resolution (MCR) aims to extract information about the underlying pure components from a series of spectra recorded for a multicomponent system [16, 17, 24]. For example, in chemical reaction systems, large amounts of data are typically available, but overlapping peaks often prevent a clear decomposition into pure component spectra and concentration profiles. MCR methods apply techniques from linear algebra and optimization to uncover this hidden structure. The underlying model is the Lambert-Beer law in matrix form, which approximates the absorption matrix $D \in \mathbb{R}^{k\times n}$ as the product of a concentration matrix $C \in \mathbb{R}^{k\times s}$ and a matrix of pure component spectra $S^T \in \mathbb{R}^{s\times n}$, thus

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

Here, k is the number of mixed spectra, n is the number of frequency channels, and s is the number of pure chemical species. Reconstructing C and S from D is a nonnegative matrix factorization (NMF) problem [5]. Two major challenges arise when applying MCR to experimental data. First, finding a solution that satisfies all imposed constraints [11, 14], and second, resolving the ambiguity of the solution – the question of which factorization is the correct one [21]. This ambiguity is often referred to as *rotational ambiguity* [2, 6, 13, 31]. In this paper, we focus on the second issue, which has been extensively studied using various approaches.

Prominent strategies include the signal contribution function [4, 34], Borgen-Rajko plots [2, 23], triangle enclosures [7], polygon inflation [27, 28], ray-casting [30], and the sensor-wise NBANDS method [18, 20]. The most recent method is the NMF-sampling algorithm by Laursen and Hobolth [12], which explores ambiguity by iteratively computing many NMFs or MCR solutions. This algorithm works for any number of chemical species and is similar to the particle swarm approach in [32]. It is both simple and effective. In [12], the algorithm is applied to cancer data, where noise and negative entries are negligible and thus not addressed.

For spectroscopic data, however, inappropriate handling of noise and small negative entries can distort the analysis. While this may not severely affect the computation of a single factorization, it often leads to unrealistic estimates when analyzing ambiguity – either generating overly large sets of feasible solutions and feasible bands or overly sharp constraints on fingerprint regions. In such cases, reliable interpretation becomes impossible.

In general, for noisy data, it is often not possible to satisfy both $CS^T = D$ and $C, S \geq 0$ simultaneously, so these constraints must be relaxed. Even when seeking a best approximation of D, one of these conditions usually remains unmet. Existing MCR methods for noisy data either tolerate small negative entries in the factors or accept deviations from the best rank-s approximation $CS^T \approx D$, but none allow both. The same limitation applies to ambiguity analysis methods.

In this paper, we extend the sampling algorithm to relax the non-negativity constraints, enabling its effective use on noisy spectroscopic data. Although this extension increases computation cost, the algorithm remains faster than any existing method for systems with four or more components.

The paper is organized as follows: Section 2 briefly introduces applied techniques for factor reconstruction and ambiguity analysis. Section 3 describes the original sampling algorithm for noise-free data, and Section 4 presents our extension to accommodate small negative values. Section 5 reports numerical results for experimental catalytic data.

2. Factor computation and analysis by the SVD

The bilinear model (1) allows to apply linear algebra techniques to reconstruct the factors. Many MCR approaches use the singular value decomposition (SVD), which factorizes D into $U\Sigma V^T$ with $U \in \mathbb{R}^{k\times k}$ and $V \in \mathbb{R}^{n\times n}$ being orthonormal matrices and $\Sigma \in \mathbb{R}^{k\times n}$ being a zero-filled diagonal matrix with the singular values σ_i in decreasing order on its diagonal. This extracts the chemically meaningful information of D. The singular values indicate the number of components, s, as for noise-free data without rank-deficiency holds $\sigma_s > 0$, but $\sigma_{s+1} = 0$.

2.1. Factor reconstruction using the SVD

The matrices U and V contain negative entries and the SVD itself is not an MCR-solution, except for the special case $s = 1$. If U and V are reduced to their first s columns and Σ is reduced to its first s columns and rows, then this truncated SVD is a rank-s best approximation of the data [33] that forms a basis for the columns of C and S. Hence, in a second step a transformation

$$C = U\Sigma T^{-1}, \quad S^T = TV^T \tag{2}$$

is applied and a proper regular $T \in \mathbb{R}^{s\times s}$ results in an MCR-solution [16, 17]. In this case, only T needs to be computed instead of the full factors C and S. This can be done by solving a proper minimization problem with penalty terms for negative entries and singular matrices T.

2.2. Ambiguity analysis using the SVD

The ambiguity analysis can be done directly profile based [18, 20, 34] or indirectly using the SVD in a low-dimensional representation. The indirect approach results in the so-called area of feasible solutions (AFS). This approach is more complicated and outside the scope of this work. For more details, see the references [6, 7, 31]. Here, we only provide a brief introduction. Without considering the scaling of the factors and their order, we assume that T from (2) is of the type where all entries in its first column equal 1 and we focus only on the first row of T. Thus T has the form

$$T = \begin{pmatrix} 1 & x^T \\ \mathbf{1} & W \end{pmatrix} \tag{3}$$

with $x \in \mathbb{R}^{s-1}$, $W \in \mathbb{R}^{s-1\times s-1}$ and $\mathbf{1} = (1,\dots,1)^T \in \mathbb{R}^{s-1}$. The spectral AFS is the set of all x so that a W exists that result in a regular T and nonnegative profiles in C and S, thus

$$\mathcal{M}_S = \{x : \text{ exists } W \text{ so that } \operatorname{rank}(T) = s,\ C, S \geq 0\}.$$

The step from the AFS to the profiles is simple. The elements x of $\mathcal{M}_S$ are low-dimensional representations of nonnegative profiles disregarding a chemical meaning. For an $x \in \mathcal{M}_S$, the associated profile is just $a = (1, x^T)V^T$. The concentrational AFS, $\mathcal{M}_C$, is defined analogously, but it also includes the singular values. For an element $y \in \mathcal{M}_C$ the associated profile is $c = U\Sigma(1, y^T)^T$.

In the context of the AFS, two important sets are the outer polytope $\mathcal{F}$ and the inner polytope $\mathcal{I}$, see [2, 10, 23]. The outer polytope $\mathcal{F} = \{x \in \mathbb{R}^{s-1} : (1, x^T)V^T \geq 0\}$ is a superset of $\mathcal{M}_S$ with respect to a nonnegative profile in the first column of S if we apply the reconstruction given by (2) and (3). The inner polytope is the convex hull of the low-dimensional representations of all rows of D, see [31]. Combining both sets allows for a geometric interpretation of the factor reconstruction as follows: An $x \in \mathbb{R}^{s-1}$ is feasible if and only if it is in the outer polytope $\mathcal{F}$ and there exist $s - 1$ further points in the outer polytope such that the simplex spanned by these points includes the inner polytope [2, 10, 23]. The low-dimensional representation of S is in the V-space and the representation of C is in the U-space. A simplex represents a complete factorization and vice versa. The vertices in V-space represent the profiles in S and the facets represent the profiles in C. If the vertices are in the outer polytope, applies $S \geq 0$ and if the facets do not intersect the inner polytope, applies $C \geq 0$. In Sec. 4.4, we will use the outer and the inner polytope and the geometric relationship for an explanation.

3. A sampling algorithm to approximate the set of feasible solutions

In [12], Laursen and Hobolth published a simple but efficient algorithm for approximating the set of feasible solutions. This algorithm works for any number of components s. The sampling algorithm starts with an initial nonnegative factorization, $D = CS^T$, and constructs a sequence of nonnegative matrix factorizations of D by changing one column of each factor per iteration. Since many factorizations are calculated, the algorithm provides a precise approximation of the feasible bands for the factorization problem. In [12] the transformation from one factorization to the next one is applied to factor C. Due to the typical focus on factor S in chemometrics, we do not adopt the formulas directly, but rather apply everything to the other factor. The idea remains unchanged.

The algorithm essentially uses two nested loops. In the outer loop, the algorithm modifies all s columns of factor S and thereby also the entire factor C. In the inner loop, an index runs from 1 to s and the algorithm changes all columns once. (4).

3.1. Changing one column in factor S

For each iteration of the inner loop, the algorithm modifies one column of S. We consider the change of the ith column in S. The new factor S' equals S for all columns, except column i. We select randomly $j \in \{1, \ldots, s\} \setminus \{i\}$ and compute the new column $S'(:, i)$ as affine linear combination of $S(:, i)$ and $S(:, j)$. Thus, for column i holds

$$S'(:, i) = (1 - \lambda)S(:, i) + \lambda S(:, j). \tag{4}$$

Let C' be the new concentration factor. With respect to nonnegative factors C' and S', the variable λ has to be in an interval $[\lambda_{\min}, \lambda_{\max}]$. For noise-free data, it is possible to compute the limits directly [12]. After computing the interval, a certain $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ is selected randomly by a β-distribution.

In the inner loop, i ranges from 1 to s. For each i, the algorithm computes three possible profiles for the column $S(:, i)$, namely the two profiles associated to $\lambda_{\min}$ and $\lambda_{\max}$ as well as the new profile $S'(:, i)$. Depending on the data and the current simplex some profiles may coincide. For a complete execution of the inner loop, there are $3s$ profiles in total.

The change from S to S' is in matrix notation

$$C' = C(M_{ij})^{-T}, \qquad S' = SM_{ij}$$

with the matrix $M_{ij} \in \mathbb{R}^{s \times s}$ defined elementwise as

$$M_{ij}(\lambda)\big|_{uv} = \begin{cases} 1 - \lambda & \text{if } u = v = i, \\ 1 & \text{if } u = v \neq i, \\ \lambda & \text{if } u = j,\ v = i, \\ 0 & \text{ow.} \end{cases} \tag{5}$$

and $(M_{ij}(\lambda))^{-1}$ being elementwise

$$M_{ij}(\lambda)^{-1}\big|_{uv} = \begin{cases} \frac{1}{1-\lambda} & \text{if } u = v = i, \\ 1 & \text{if } u = v \neq i, \\ \frac{\lambda}{\lambda-1} & \text{if } u = j,\ v = i, \\ 0 & \text{ow.} \end{cases}$$

3.2. Boundaries for the variable λ

To ensure nonnegativity for the new factors the variable λ has to be in $[\lambda_{\min}, \lambda_{\max}]$ with

$$\lambda_{\min} = \max_{\ell \in I} \frac{S_{\ell i}}{S_{\ell i} - S_{\ell j}}, \quad \lambda_{\max} = \min_{\ell \in I'} \frac{C_{\ell j}}{C_{\ell i} + C_{\ell j}} \tag{6}$$

for $I = \{\ell : S_{\ell j} > S_{\ell i}\}$ and $I' = \{\ell : C_{\ell i} + C_{\ell j} > 0\}$. The interpretation of this change in the low-dimensional representation with the SVD is as follows: In the V-space the vertex associated with $S(:, i)$ of the representing simplex shifts in the direction of the vertex associated with $S(:, j)$. The inner and the outer polytope limit the variable λ. The inner polytope specifies $\lambda_{\max}$ and the outer polytope specifies $\lambda_{\min}$. Fig.1 visualizes this geometric interpretation for an example with $s = 3$. The transformation $S' = SM_{ij}$ can also be applied directly to the low-dimensional representation. Let $S^T = TV^T$ with ones in the first column of T. Then $S' = SM_{ij} = VT^TM_{ij} = V\tilde{T}^T$ and $\tilde{T}$ has only ones in its first column.

3.3. Additional profiles for the dual factor

The change of column i in S as in (4) changes only column j in C. Thus the computation of three feasible columns for $S(:, i)$ results in three profiles for $C(:, j)$. For one step of the outer loop i goes from 1 to s and the algorithm computes a total of $3s$ new profiles. For factor C the total number of computed nonnegative profiles is the same but as j is selected randomly with $j \neq i$ per iteration there are not necessarily three new profiles per column.

3.4. Stopping criterion

Let $C^{(i)}, S^{(i)},\ i = 0, 1, 2, \ldots$ be the factors of the computed feasible factorizations. We collected the factors in the sets

$$\mathcal{C}^m = \{C^{(0)}, C^{(1)}, \ldots, C^{(m)}\},\ \mathcal{S}^m = \{S^{(0)}, S^{(1)}, \ldots, S^{(m)}\}$$

and define, with m being the number of computed factors, a measure for the range of feasible profiles in C as

$$\text{avg}\langle \mathcal{C}^m \rangle = \frac{1}{ks} \sum_{i=1}^{k} \sum_{j=1}^{s} \left(\max_{\ell=0,\ldots,m} C_{ij}^{(\ell)} - \min_{\ell=0,\ldots,m} C_{ij}^{(\ell)} \right).$$

In [12] Laursen and Hobolth suggest to use the change in the range of profiles in factor C for the stopping criterion. Every $\tau = 1000$ iterations of the outer loop a check is made whether

$$\text{avg}\langle \mathcal{C}^m \rangle - \text{avg}\langle \mathcal{C}^{m-\tau} \rangle < \epsilon$$

applies, where $\epsilon > 0$ is a suitable control parameter. For spectroscopic data the stopping condition

$$\frac{\text{avg}\langle \mathcal{C}^m \rangle - \text{avg}\langle \mathcal{C}^{m-\tau} \rangle}{\text{avg}\langle \mathcal{C}^{m-\tau} \rangle} + \frac{\text{avg}\langle \mathcal{S}^m \rangle - \text{avg}\langle \mathcal{S}^{m-\tau} \rangle}{\text{avg}\langle \mathcal{S}^{m-\tau} \rangle} < \epsilon$$

might be better.

3.5. No skipping between AFS segments

The AFS can be a connected subset with a whole around the origin, or it can consist of several subsets (segments). In the case of separated subsets, the question is whether it is possible that the vertices of the simplex can change between the segments. Due to the algorithm's construction, that only one column of S is modified per iteration in the inner loop, the new profile remains in the same subset of the AFS. This is in contrast to other methods based on the signal contribution function [34] or the approach of sensorwise N-bands. In these methods, the transformation from one factor to another is done by a regular matrix $T \in \mathbb{R}^{s \times s}$ without the special structure of M_{ij} from the sampling algorithm. This means that the algorithm is stable in this respect and does not require additional control or subsequent sorting of the points. This benefits the implementation.

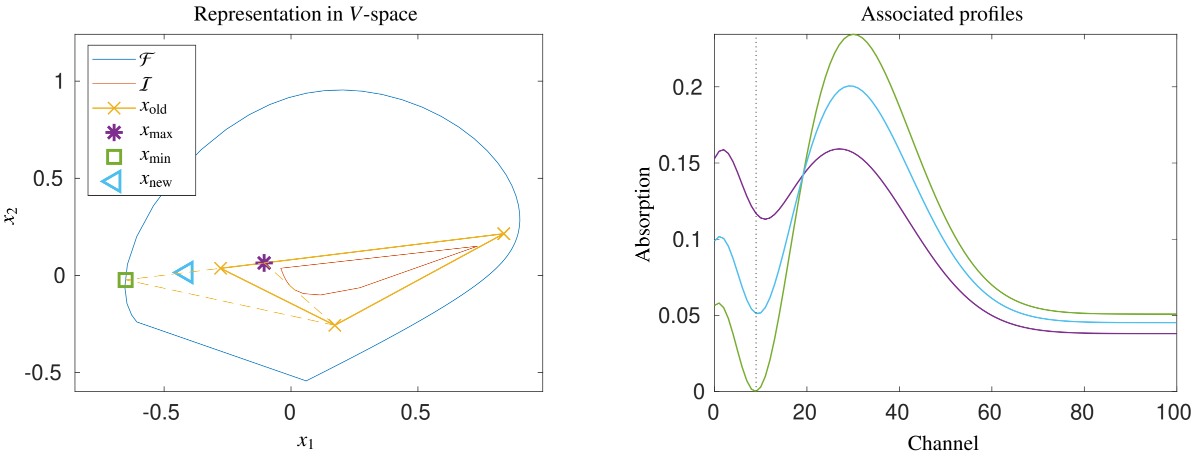

Figure 1: Visualization of the sampling algorithm in the low-dimensional representation for a model problem from FACPACK. Left: The current S defines a simplex (simplex with solid ochre lines). One vertex of the simplex is changed in the direction of another vertex by the variable λ as in (4). Nonnegativity constraints for C and S limit λ. If the vertex of $S(:, i)$ is changed by $S(:, j)$ with $i \neq j$, then the condition $S'(:, i) \geq 0$ defines the lower limit for λ namely λ_{min} and the condition $C'(:, j) \geq 0$ defines λ_{max} with S' and C' being the modified factors. In the left plot the points associated with λ_{min} and λ_{max} are labeled as x_{min} and x_{max}. Right: the profiles associated to the points in the left graphic. The black dotted lines marks the index associated with the limiting line of $\mathcal{F}$ in the left graphic.

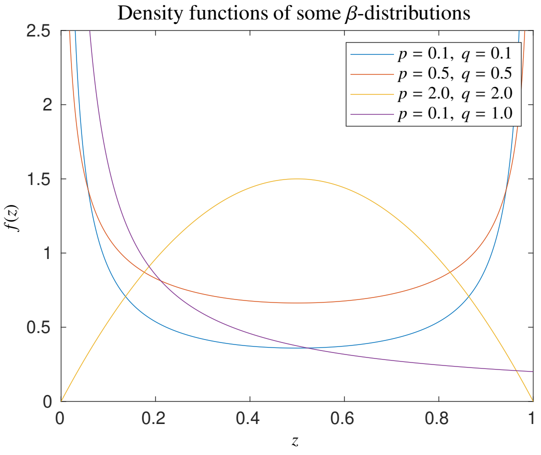

Figure 2: The density function of the β-distribution for different parameters $p, q > 0$.

3.6. *Selection of the new profile using the β-distribution*

The new column $S'(:, i)$ is computed as in (4) by the variable λ using the β-distribution. The β-distribution is defined on the interval $[0, 1]$ and has the density function

$$f(z) = \frac{1}{B(p, q)} z^{p-1}(1 - z)^{q-1}$$

with control parameters $p, q > 0$ as well as $B(p, q) = \int_0^1 u^{p-1}(1 - u)^{q-1}\mathrm{d}u$. The parameter p mainly controls the behavior in the left half, between 0 and 0.5, while q mainly controls the behavior in the right half, between 0.5 and 1. Small values of p and q, e.g. $p, q < 1$, favor values close to the boundaries 0 and 1 rather than close to 0.5. See Fig. 2 for some selections of p and q.

Small values of p, q result in a good sampling progress, because the new columns are close to one of the extreme profiles and the associated points in the low-dimensional representation are close to the boundaries of the AFS. The same number of computed factors results in a better approximation of the set of possible profiles. Conversely, the algorithm achieves a good approximation in fewer steps, and the stopping criterion is met within less iterations, see Fig. 3 for a comparison.

4. NMF sampling for noisy data

Noise is often an unavoidable ingredient in the analysis of experimental data. The sampling algorithm can easily be adapted for use with noisy data. Before introducing the necessary modifications, we briefly discuss how to compute meaningful approximations to the set of feasible profiles in the presence of noise.

Depending on the spectroscopic technique and the experiment itself, noise can cause various difficulties. For example, noise can generate negative entries in the data due to a background subtraction. It can also prevent a precise approximation of the data by a low-rank factorization, as well as a factorization or best approximation by nonnegative factors. Figure 4 shows the difficulties that can arise when computing a factorization in the low-dimensional representation for spectroscopic data, even though the data has a very high signal-to-noise ratio and a good low-rank approximation.

4.1. *Options to handle noise*

Computing the range of feasible profiles respectively the area of feasible solutions for noisy data is challenging and different approaches are available [6, 7, 19, 20, 27]. Three options are:

1. Compute initial factors C and S^T that result in a best approximation or at least in a small error $\|D -$

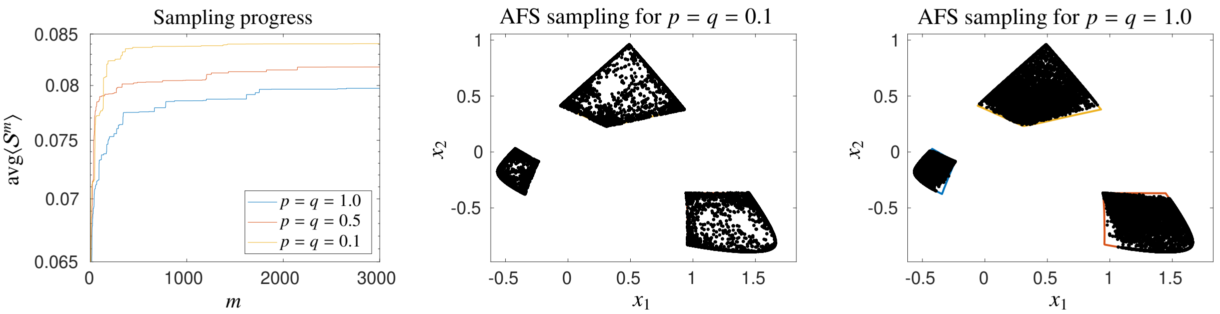

Figure 3: Application of the sampling algorithm to a three component data set from FACPACK with $k = 21$ and $n = 101$ for different control parameters of the β-distribution. The algorithm is applied with 3000 iterates of the outer loop. Left: the progress using the measure avg$\langle\mathcal{S}^m\rangle$ with $\mathcal{S}^m$ being the set of feasible profiles for S with the first m iterations of the outer loop. Center and right: low-dimensional representation of the sampling results after 3000 iterations of the outer loop. Higher values for p, q result in more points in the interior of the AFS segments and values close to 0 in more points close to the boundaries. Thus the progress is better for small values of p, q.

$CS^T\|$, set all negative entries in the factors to 0 and work with the new product as

$$\tilde{D} = C_+S_+^T \text{ with } C_+ = \max(C, 0),\ \ S_+ = \max(S, 0).$$

By construction, the matrix $\tilde{D}$ has no negative entries and a factorization with nonnegative factors of rank s. However, the problem with this approach is the strong dependence on the initial factors for the further analysis. If either one profile in C or S has major negative entries in at least one window including a fingerprint, then the new product, $\tilde{D}$, is significantly changed in this area and the subsequent ambiguity analysis may be meaningless.

2. Compute a proper initial factorization as in approach 1 but use the error

$$\epsilon := \left\|D - C_+S_+^T\right\|^2$$

as a bound/reference to classify other factorizations as feasible or not. So a computed factorization by C' and S' is classified as feasible, if

$$\left\|D - (C')_+(S'_+)^T\right\|^2 < (1 + \delta)\epsilon$$

with an initially defined small control parameter $\delta > 0$. Typically the Frobenius-Norm $\|A\|_F = \text{ssq}(A_{ij})^2$ is used for this application [8, 17]. A disadvantage of this approach is that the Frobenius norm of a $k \times n$-matrix must be computed for each factorization. This is computationally expensive.

3. Basically different to the first two approaches is the idea to allow small deviation from the nonnegativity constraints in the form that a factorization is classified as feasible if for all $\ell = 1, \ldots, s$ holds

$$\frac{\min_i(C_{i\ell})}{\max_i(C_{i\ell})} \geq -\varepsilon, \qquad \frac{\min_j(S_{j\ell})}{\max_j(Y_{j\ell})} \geq -\varepsilon \tag{7}$$

with a small control parameter $\varepsilon \geq 0$. This approach is as flexible as approach 2, but it has the following advantage: If we compute C and S with the help of an SVD or as $C = C^{(\text{old})}M^{-1}$ and $S = S^{(\text{old})}M^T$ with a regular matrix M, then this transformation does not change the error, so it is not necessary to compute it for each iteration.

4.2. Extension of the sampling algorithm for noisy data

Next, we extend the NMF-sampling to noisy data, focusing on spectroscopic data with sharp peaks and possibly several non-absorbing areas. Approach 1 has the advantage that the algorithm requires no adaptation, but the results depend strongly on the initial factorization and on whether (wrongly computed) signals in negative direction are cut. The disadvantage of approach 2 is its high computational cost to evaluate whether a profile is feasible or not. In this paper, we use the approach 3, constructing the profiles with a truncated SVD and accepting relatively small negative entries in the profiles.

The main modification regarding the sampling algorithm and approach 3 for working with noisy data is that we cannot compute the boundaries $\lambda_{\min}$ and $\lambda_{\max}$ as in (6). These computations focus on strictly nonnegative profiles, which may not even exist for noisy data and a reconstruction via the SVD. For accepting relatively small negative entries as in (7) we must approximate $\lambda_{\min}$ and $\lambda_{\max}$ using the bisection method. Although the computational costs are much higher than using the explicit computation for noise-free data, they are still acceptable. Algorithm 1 shows a pseudo-code element for the sampling.

4.3. Quick computation of the limits for λ

Next, we explain the decisive difference to the original sampling. The interval for λ is slightly extended so that the condition from (7) holds. Note that when computing $\lambda_{\min}$ and $\lambda_{\max}$, we do not consider the complete factors. Due to the special construction of $M_{ij}(\lambda)$, only one column of C respectively only one column in S is checked, as only one column is modified. In particular with $M_{ij}(\lambda)$ as in (5) the only non-rescaling change from S to S' is in $S'(:, i)$ and the only change from C to C' is in $C'(:, j)$. Thus, only these columns limit the variable λ.

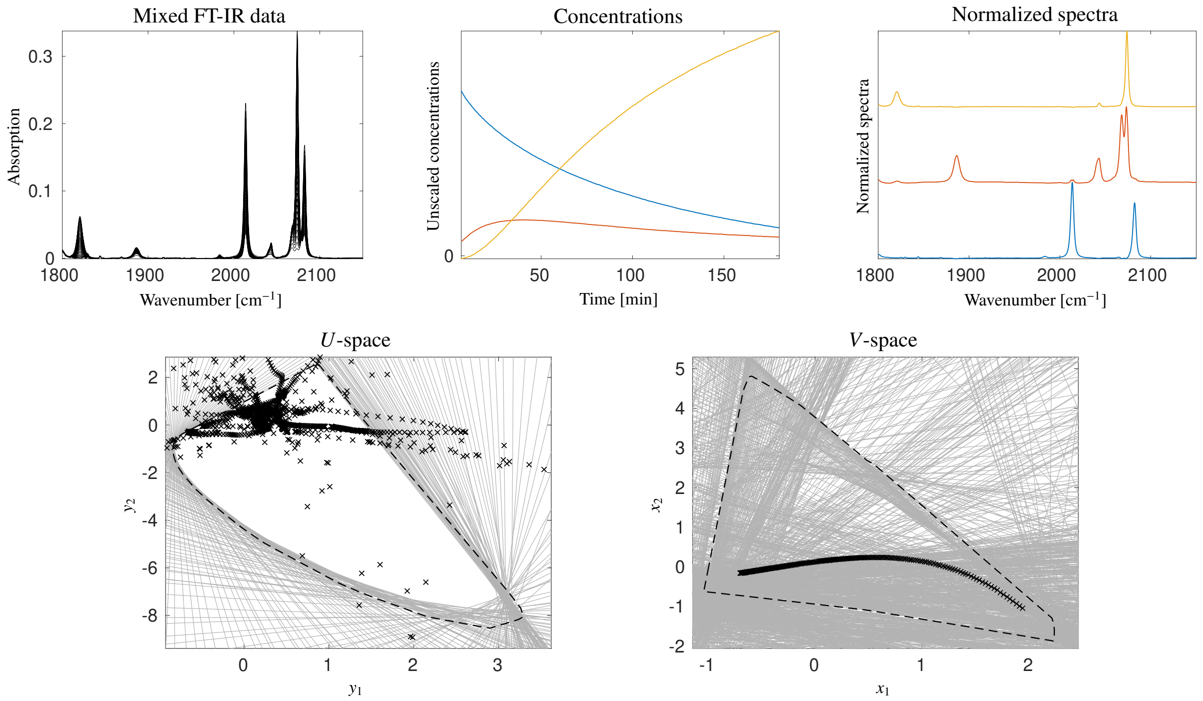

Figure 4: The in situ FT-IR data from a mixture of rhodium carbonyl complexes reveals the problem of even small perturbations when computing a factorization with rigorously nonnegative constraints. Upper plots: the data as well as a factorization with Rh(acac)(CO)$_2$ (blue), Rh$_4$(CO)$_{12}$ (red) and Rh$_6$(CO)$_{16}$ (yellow). The data has a very high signal-to-noise ratio and the factorization is a very good approximation of the data. Lower plots: nevertheless a nonnegative factorization based on the first 3 singular values and vectors is not possible. The gray lines represent the nonnegativity constraints and the black crosses are the data points. No triangle exists that is in the outer polygon and includes all data points. The black dashed-lines are approximations to the outer polygons with respect to relatively small negative entries in the columns of C respectively S not larger than 0.01.

The restriction $\min(C'(:,j))/\max(C'(:,j)) \geq -\varepsilon$ defines $\lambda_{\max}$ and the restriction $\min(S'(:,i))/\max(S'(:,i)) \geq -\varepsilon$ defines $\lambda_{\min}$. Theoretically, to approximate $\lambda_{\max}$, we must also check whether the factorization degenerates, meaning the associated simplex in the low-dimensional representation excludes the inner polytope completely. These are rare cases and not relevant in practice. The bisection method with small steps prevents this scenario. The rest of the algorithm can remain unchanged. In Algorithm 1 this situation is handled by checking whether in the low-dimensional representation the new simplex still includes the origin.

4.4. The inner polytope as a byproduct

The approximation of the inner polytopes, $\mathcal{I}_S$ and $\mathcal{I}_C$, is very simple for noise-free data. They are the convex hulls of

$$a_i = \frac{(U\Sigma)^T(2:s,i)}{U_{i1}\sigma_1},\ i = 1\ldots,k,$$

respectively

$$b_j = \frac{V^T(2:s,j)}{V_{j1}},\ j = 1,\ldots,n.$$

With the help of the inner polytopes it is possible to compute the outer polytopes using duality [9, 22, 25]. For example, the vertices of $\mathcal{F}_S$ are dual to the facets of $\mathcal{I}_C$.

The situation is fundamentally different for noisy data. If we consider relatively small negative entries in the factors as in (7), one can approximate the outer polytope numerically using inverse polygon inflation [29] or ray casting [30], but with large effort for $s = 4$ and huge effort for $s > 4$. Subsequently, duality enables the computation of approximations of the inner polytopes. Using the sampling algorithm, one can compute a very good approximation of the inner polytopes as a byproduct. Let $\tilde{\mathcal{S}}$ be the set of all computed simplexes for factor S in low-dimensional representation, and let $\mathcal{I}_S$ be the inner polygon in V-space. It holds

$$\mathcal{I}_S \supseteq \bigcap_{\tilde{S}\in\tilde{\mathcal{S}}} \tilde{S}$$

and since $\tilde{\mathcal{S}}$ contains a wide range of possible factors the intersection of all $\tilde{S} \in \tilde{\mathcal{S}}$ is a good approximation to $\mathcal{I}_S$ with respect to small negative entries in S.

4.5. Finding AFS from the sampling results

Obtaining the AFS from point-set that result from the sampling algorithm is not trivial. The main difficulty is that the isolated AFS subsets are generally not convex, so a convex hull would result in an AFS subset approximation that is too large. The inner boundary consists of segments of conic sections, see [1], and it is more difficult to approximate this shape.

```
Set $m := 1$
while $m < maxiter$ do
    for $i = 1 : s$ do
        Select $j \in \{1, \dots, s\}$ randomly with $j \neq i$
        Approximate $\lambda_{\max}$ using bisection as the
          maximum so that $C' = C(M_{ij}(\lambda))^{-1}$ fulfills
          $\frac{\min C'(:,j)}{\max C'(:,j)} \geq -\varepsilon$ and in the low-dimensional
          representation the new simplex still includes
          the origin
        Approximate $\lambda_{\min}$ as the minimum so that
          $S' = S\,M_{ij}(\lambda)$ fulfills $\frac{\min S'(:,i)}{\max S'(:,i)} \geq -\varepsilon$ using
          bisection
        Select $\lambda \in [\lambda_{\min}, \lambda_{\max}]$ using the $\beta$-distribution
        Add the new columns to the list of possible
          columns for $C$ and $S$
        Add the new points in $U$- and $V$-space to the
          list of feasible points
        Continue with $S := S'$ and $C := C'$
    end
    if $mod(m, M) = 0$ then
        Check the stopping criterion
    end
    Set $m := m + 1$
end
```

Algorithm 1: Pseudo-code element for the sampling algorithm with respect to small negative entries in C and S.

We propose two approaches to approximate the AFS using the results of the sampling algorithm. First, the results of the algorithm provide inner and outer polytopes as byproducts. Then analytical approaches can be used to calculate the inner boundary curve and the AFS from the inner and outer polytopes, as discussed in [1, 23].

Alternatively, the AFS can be calculated directly from the sampling results using an α-shape, a generalization of the convex hull. The concept of an α-shape of a finite set of points was introduced by Edelsbrunner et al. [3] and is a common tool for shape reconstruction in computational geometry. In this work, we use the Matlab function `alphaShape`, which helps to reconstruct the AFS shape from the low-dimensional representation of the sampled profiles. The α-shape depends on the chosen α value. Fig. 5 shows the results for a set of points from one of the subsets of the concentration AFS in Fig. 6. An optimal value of α seems to be five times the smallest value of α value that produces an α-shape with only one region.

5. Numerical results

5.1. Application to in situ FTIR spectroscopic data with three components

First, we apply the sampling algorithm to in situ FTIR spectroscopic data from a P-ligand free rhodium catalysis. In the process, $Rh(acac)(CO)_2$ was progressively transformed to $Rh_4(CO)_{12}$ and $Rh_6(CO)_{16}$ under typical hydroformylation conditions: 100°C, 20 bar of CO/H_2, CO/H_2=1:1, with dodecane as solvent and an initial $Rh(acac)(CO)_2$ concentration of $1 \cdot 10^{-3}$ mol L^{-1}, see [15] for more details. Although the system has a high signal-to-noise ratio, it contains small negative entries. Additionally, there is also an asymmetrical distribution of negative entries in U- and V-space as

$$\min\left(\frac{\min D(:,i)}{\max D(:,i)}\right) = -86.5,$$
$$\min\left(\frac{\min D(i,:)}{\max D(i,:)}\right) = -4.41 \cdot 10^{-3}.$$

The mixed data as well as the pure factors are shown in Fig. 4. The data contains $k = 208$ spectra with $n = 1453$ wavenumbers in the range [1800, 2150]cm^{-1} and absorptions from the $s = 3$ species $Rh(acac)(CO)_2$, $Rh_4(CO)_{12}$ and $Rh_6(CO)_{16}$.

The control parameters are:

- $\varepsilon = 0.01$, to accept relatively small negative entries in C and S,
- $p = q = 0.05$ for the β-distribution,
- $\epsilon = 10^{-4}$ and $maxiter = 4000$ for the stopping criterion of the sampling.

The algorithm stops after 3000 iterations of the outer loop. Thus, it computes 9000 profiles per column of S. The numbers of computed columns of the dual factor C are slightly different. In this cases the algorithm computes 9036 feasible profiles for $C(:,1)$, 9051 for $C(:,2)$, and 8913 for $C(:,3)$. The numbers differ because the affected dual factor depends on the randomly selected $j \neq i$, with $j \in \{1, \dots, s\}$. The computation took 13.9 s on an Intel i7 CPU with 2.9 GHz, 8 cores, and 64 GB of RAM. The algorithm was executed entirely as a Matlab-file. Computing both AFS sets with polygon inflation implemented in C took 9.22 s. The results for the AFS sets, as well as the band boundaries for the profiles, are shown in Fig. 6.

5.2. Application to a model problem with an AFS with a punctiform and a straight

The sampling algorithm works also stable for degenerated data. For the following example with $k = 101$, $n = 51$, and $s = 3$, the spectral and the concentration AFS consist of an isolated point, a straight, and a bounded area. We apply the algorithm with $\varepsilon = 2 \cdot 10^{-7}$ and detect the AFS. In the concentration AFS, the straight is a narrow area, due to small deviations from the strict condition $C, S \geq 0$. The results are presented in Fig. 7.

5.3. Application to in situ FTIR spectroscopic data with four components

Finally, we apply the algorithm to in situ FTIR spectroscopic FTIR data from a sequential dosage experiment. The experiment started with a solution of $Rh(acac)(CO)_2$ ($2.5 \cdot 10^{-3}$ mol L^{-1} in Me-THF) under inert conditions (80° C, 1 bar Ar). Upon introduction of 20 bar of CO/H_2 (1:1), $Rh(acac)(CO)_2$ was constantly converted to $Rh_6(CO)_{16}$. After reaching a steady state,

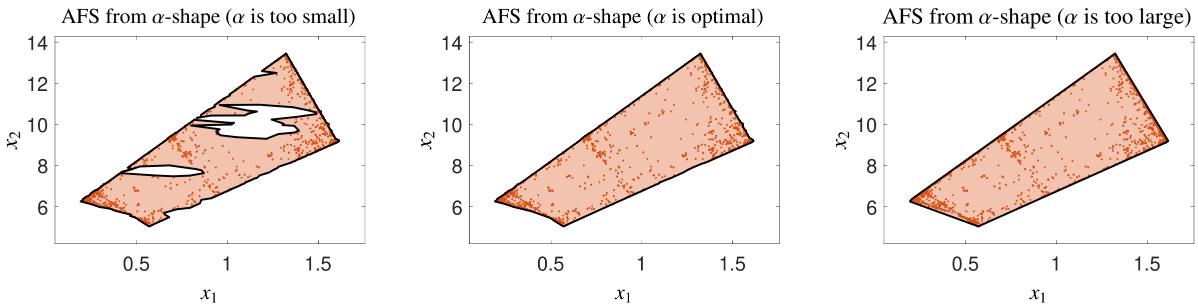

Figure 5: The α-shapes for one AFS segment in U-space for the results presented in Fig. 6. The α value is too small in the left plot ($\alpha = 0.13585$), optimal in the middle plot ($\alpha = 0.67924$) and too large in the right plot ($\alpha = \infty$), where the α-shape is the convex hull.

the Biphephos ligand (with intended concentration of $5.0 \cdot 10^{-3}$ mol L^{-1}) was added, leading to the transformation of $Rh_6(CO)_{16}$ to $HRh(CO)_2(PP)$ complex. The data contains $k = 979$ mixed spectra. In each spectrum we only consider the interval $[1900, 2300]$cm^{-1} with $n = 1660$ wavenumbers. In this range, $s = 4$ species have absorptions: the rhodium complexes $Rh(acac)(CO)_2$, $Rh_6(CO)_{16}$, $HRh(CO)_2(PP)$ and CO. The original data contains also information from other species but, in order to demonstrate the effectiveness of the algorithm, we select a frequency range with only four absorbing species.

First, we compute a factorization using the Automated Peak Group Analysis [26]. The data, as well as the concentration profiles and the pure component spectra are shown in Fig. 8. Next, we apply the sampling algorithm, allowing for negative entries up to an amount of $\varepsilon = 0.026$. The algorithm terminates after $m = 3000$ runs of the outer loop. Fig. 9 shows the results in U- and V-space, as well as the band boundaries for the concentrations and pure component spectra of the four species. The computation took 22.8 seconds on the same computer as used in Sec. 5.1. Computing the AFS sets by ray-casting from FACPACK using 5000 rays took 27.8min on the same computer.

6. Conclusion and outlook

For the analysis of spectroscopic data, fast computations of pure component decompositions and assessment of their ambiguity are desirable. In recent years, much attention has been given to accurately approximating the AFS. However, for rapid and exploratory analyses, especially in high-dimensional data and systems with more than three components, sampling offers an effective alternative. While the resulting AFS sets may lack precision at some vertices, these algorithms are sufficiently accurate when the random factorizations are well controlled.

The key advantage of the applied type of sampling is that it avoids computationally expensive optimization steps, which typically dominate the runtime of methods like grid search, polygon inflation, ray-casting, and the sensor-wise N-bands algorithm. The extension of the sampling method from [12] to handle noisy data adds a powerful tool to the collection of techniques for approximating AFS sets and feasible bands. Its structure is well suited to parallel computing, providing additional time savings in practice. A future integration into the FACPACK software is planned.

7. Acknowledgement

The German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) is thanked for partly funding this study within the framework of the Research Unit IMPD4Cat (FOR5538, project number: 501735683, KU 4283/1-1).

References

[1] T. Andersons, M. Sawall, and K. Neymeyr. Analytical enclosure of the set of solutions of the three-species multivariate curve resolution problem. *J. Math. Chem.*, 60:1750–1780, 2022.

[2] 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.

[3] H. Edelsbrunner, D.G. Kirkpatrick, and R. Seidel. On the shape of a set of points in the plane. *IEEE Trans. Inf. Theory*, 29(4):551–559, 1983.

[4] P. J. Gemperline. Computation of the range of feasible solutions in self-modeling curve resolution algorithms. *Anal. Chem.*, 71(23):5398–5404, 1999.

[5] N. Gillis. *Nonnegative Matrix Factorization*, volume 1 of *Data science*. SIAM, Philadelphia, 2021.

[6] 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.

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

[8] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3 of *Johns Hopkins Stud. Math. Sci.* Johns Hopkins University Press, Baltimore, 2012.

[9] R. C. Henry. Duality in multivariate receptor models. *Chemom. Intell. Lab. Syst.*, 77(1-2):59–63, 2005.

[10] 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.

[11] 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(2):713–730, 2008.

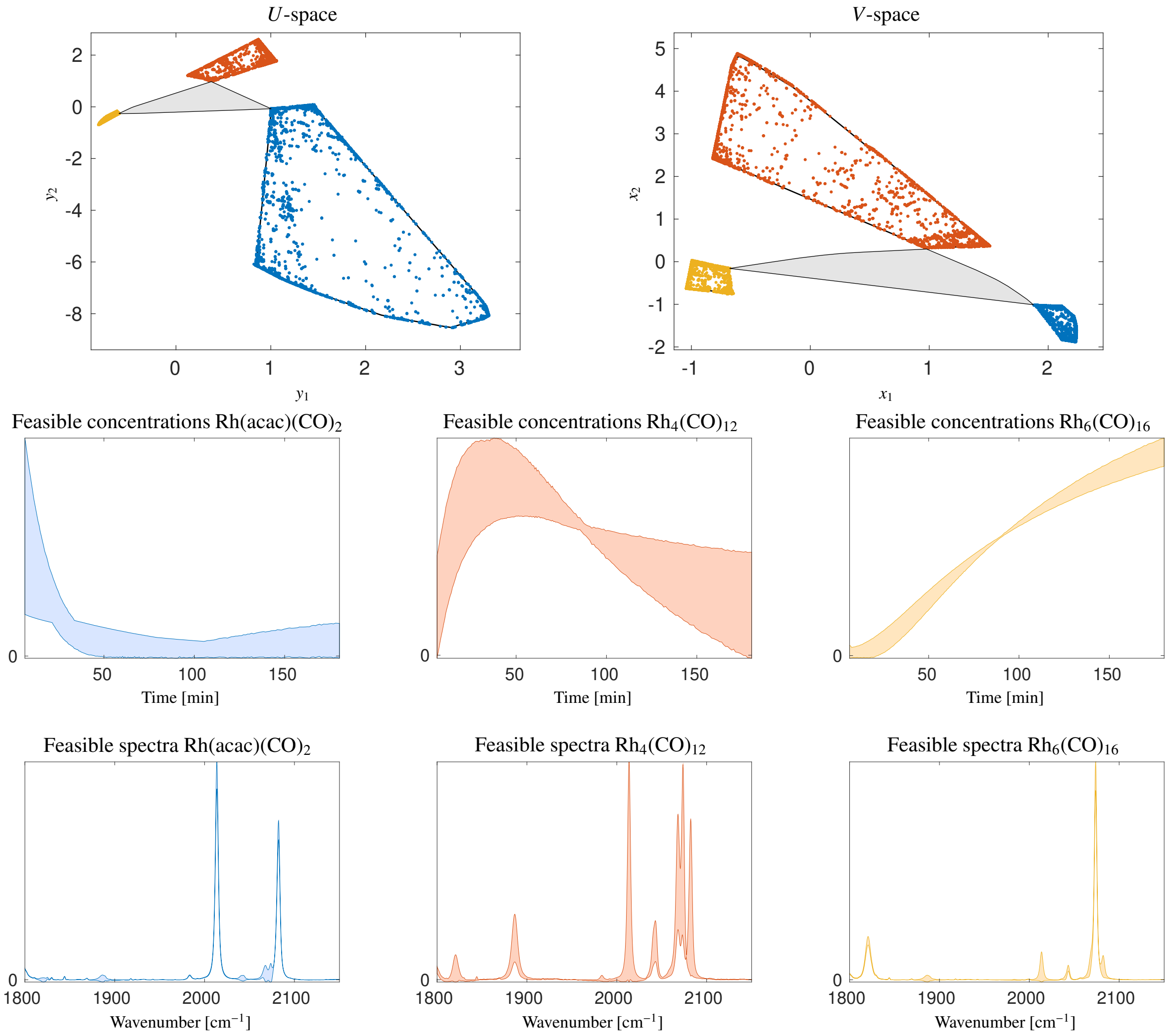

Figure 6: Analysis of the rotational ambiguity by the sampling algorithm for the data from Sec. 5.1. Top: The sampled AFS sets (colored points), the inner polytopes (gray) and the polygon inflation results (black lines) for the spectroscopic data obtained for the P-ligand free rhodium catalyst system. Center and bottom: the computed feasible band boundaries for the concentrations and the spectra.

[12] R. Laursen and A. Hobolth. A sampling algorithm to compute the set of feasible solutions for nonnegative matrix factorization with an arbitrary rank. *SIAM J Matrix Anal. Appl.*, 43(1):257–273, 2022.

[13] W. H. Lawton and E. A. Sylvestre. Self modelling curve resolution. *Technometrics*, 13(3):617–633, 1971.

[14] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. *Adv. Neural Inf. Process. Syst.*, 13:556–562, 2001.

[15] B. N. Leidecker, D. P. Fuentes, C. Wei, M. Sawall, K. Neymeyr, R. Franke, A. Börner, and C. Kubis. In situ FTIR spectroscopic investigations on rhodium carbonyl complexes in the absence of phosphorus ligands under hydroformylation conditions. *New J. Chem.*, 48:18365–18375, 2024.

[16] M. Maeder and Y. M. Neuhold. *Practical data analysis in chemistry*, volume 26 of *Data Handling Sci. Technol.* Elsevier, Amsterdam, 2007.

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

[18] A. C. Olivieri. Estimating the boundaries of the feasible profiles in the bilinear decomposition of multi-component data matrices. *Chemom. Intell. Lab. Syst.*, 216:104387, 2021.

[19] A. C. Olivieri, M. Sawall, K. Neymeyr, and R. Tauler. Noise effects on band boundaries in multivariate curve resolution of three-component systems. *Chemom. Intell. Lab. Syst.*, 228:104636, 2022.

[20] A. C. Olivieri and R. Tauler. N-BANDS: A new algorithm for estimating the extension of feasible bands in multivariate curve resolution of multicomponent systems in the presence of noise and rotational ambiguity. *J. Chemom.*, 35(3):e3317, 2021.

[21] 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.

[22] R. Rajkó. Natural duality in minimal constrained self modeling curve resolution. *J. Chemom.*, 20(3-4):164–169, 2006.

[23] 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.

[24] C. Ruckebusch and L. Blanchet. Multivariate curve resolution: A review of advanced and tailored applications and challenges. *Anal. Chim. Acta*, 765:28–36, 2013.

[25] M. Sawall, C. Fischer, D. Heller, and K. Neymeyr. Reduction of the rotational ambiguity of curve resolution techniques under partial knowledge of the factors. Complementarity and coupling theorems. *J. Chemom.*, 26(10):526–537, 2012.

[26] M. Sawall, C. Kubis, B. N. Leidecker, L. Prestin, T. Andersons, M. Beese, J. Hellwig, R. Franke, A. Börner, and K. Neymeyr. An automated Peak Group Analysis for vibrational spectra analysis. *Chemom. Intell. Lab. Syst.*, 254:105234, 2024.

[27] M. Sawall, C. Kubis, D. Selent, A. Börner, and K. Neymeyr. A

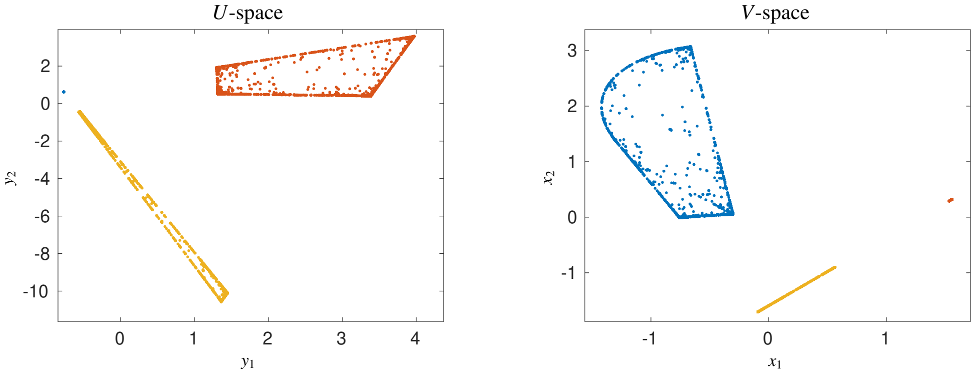

Figure 7: The results of the sampling algorithm for a data set with an isolated point , a straight and an area for the spectral AFS as well as for the concentration AFS. For the concentration AFS the straight is only a narrow area, aggravated by the different scaling of y_1 and y_2 due to the singular values.

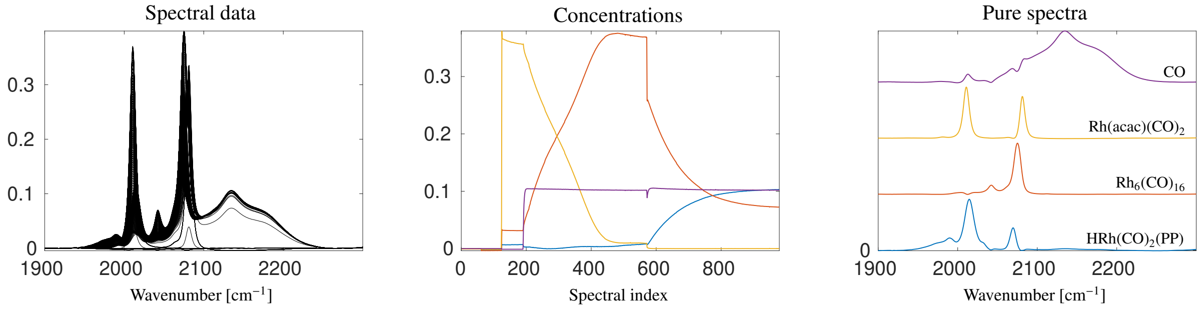

Figure 8: The in situ FT-IR spectroscopic data from Sec. 5.3. The measured spectra (right) as well as the unscaled pure factors of the true solution.

fast polygon inflation algorithm to compute the area of feasible solutions for three-component systems. I: Concepts and applications. *J. Chemom.*, 27(5):106–116, 2013.

[28] 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 FACPACK implementation. *J. Chemom.*, 28(5):633–644, 2014.

[29] 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.

[30] M. Sawall and K. Neymeyr. A ray casting method for the computation of the area of feasible solutions for multicomponent systems: Theory, applications and FACPACK-implementation. *Anal. Chim. Acta*, 960:40–52, 2017.

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

[32] A. N. Skvortsov. Estimation of rotation ambiguity in multivariate curve resolution with charged particle swarm optimization. *J. Chemom.*, 28(10):727–739, 2014.

[33] G. W. Stewart. On the early history of the singular value decomposition. *SIAM Rev.*, 35:551–556, 1993.

[34] R. Tauler. Calculation of maximum and minimum band boundaries of feasible solutions for species profiles obtained by multivariate curve resolution. *J. Chemom.*, 15(8):627–646, 2001.

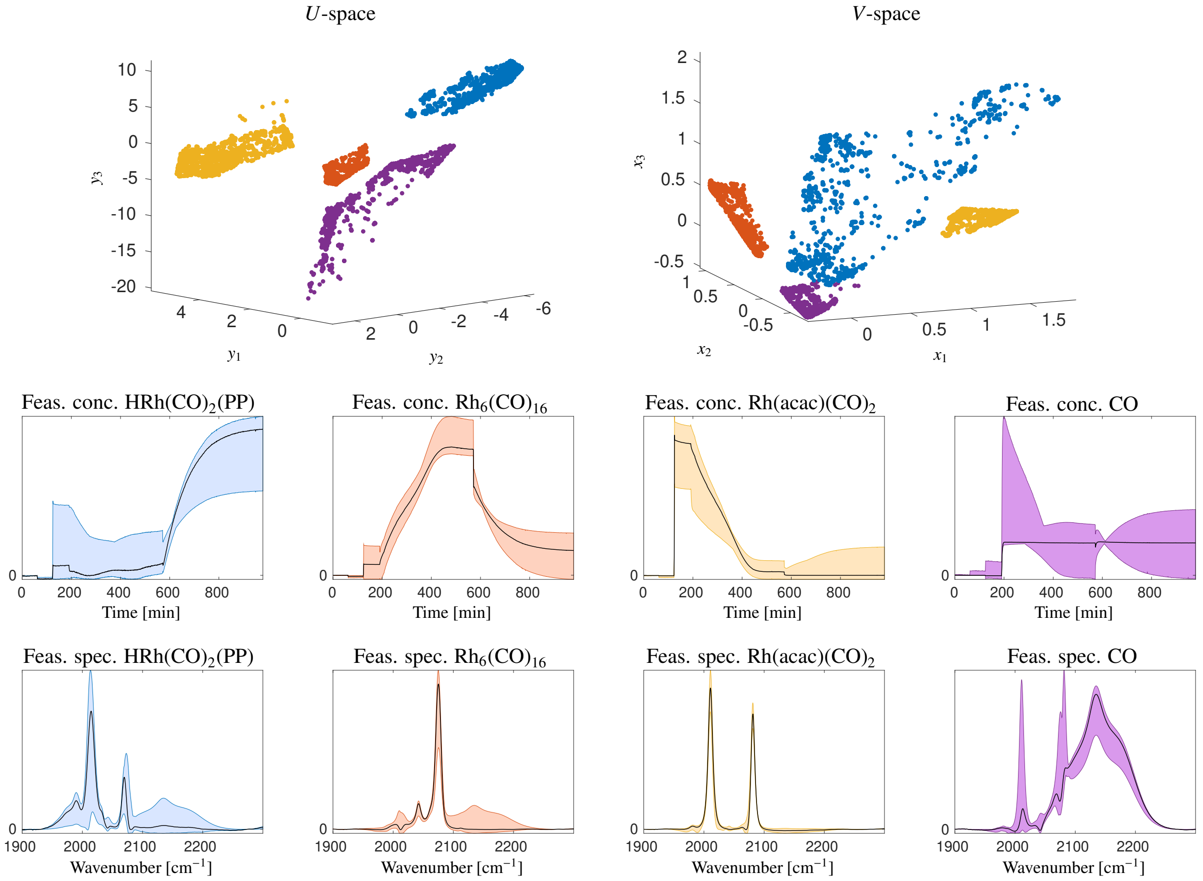

Figure 9: The approximated AFS-sets as well as the feasible band boundaries for the four-component in situ FT-IR spectroscopic data from Sec. 5.3. The AFS-sets consists of four separated subsets. A sampling with only 3000 runs of the outer loop results in a number of 9000 points per subset (ambiguities included). The band boundaries include the true profiles (black).