An active constraint approach to identify essential spectral information in noisy data

Mathias Sawall[a], Cyril Ruckebusch[b], Martina Beese[a,d], Robert Francke[c,d], Adrian Prudlik[c,d], Klaus Neymeyr[a,d]

[a] *Universität Rostock, Institut für Mathematik, Ulmenstrasse 69, 18057 Rostock, Germany*
[b] *Université de Lille, CNRS, Laboratoire de spectroscopie avancé, interactions, réactivité et environnement (LASIRE), F-59000, Lille, France*
[c] *Universität Rostock, Institut für Chemie, Albert-Einstein-Straße 3a, 18059 Rostock, Germany*
[d] *Leibniz-Institut für Katalyse, Albert-Einstein-Strasse 29a, 18059 Rostock*

Abstract

Multivariate curve resolution (MCR) methods aim at extracting pure component profiles from mixed spectral data and can be applied to high-dimensional data, e.g., from process spectroscopy or hyperspectral imaging techniques. One often observes that some parts of this data, namely certain rows and columns of the data matrix, are considered essential for MCR outcomes, while other parts are of minor importance. Some methods for determining essential data are known, but all have different disadvantages concerning the application for noisy data. This work presents a new approach on how to detect the essential information for noisy, experimental spectral data. Active nonnegativity constraints in combination with duality arguments are the key ingredients for determining essential spectra and frequency channels. The new approach is conceptually simple, computationally cheap and stable with respect to noise. The algorithm is tested for noisy experimental Raman, UV-Vis and FTIR-SEC data.

Key words: multivariate curve resolution, essential spectra and frequencies, area of feasible solutions, ray casting.

1. Introduction

Modern computerized spectrometers with high time and frequency resolution typically produce high-dimensional data sets. From this data, multivariate curve resolution (MCR) techniques can extract the underlying pure component information. This information can be spectral fingerprints characteristic of particular species or even complete pure component spectra along with their associated concentration profiles [13, 14]. For high-dimensional data, MCR methods may suffer from many signals with little or no information content. Then dimension reduction techniques are desirable that enable a separation of the essential parts of the data from the parts of lesser importance. The separation can speed up subsequent MCR analyses.

Recently, techniques have been introduced for the detection of relevant and essential parts of spectral data sets [21, 5, 18, 6]. Such techniques are applicable after preparatory steps of dimensionality reduction by the singular value decomposition (SVD, [9]) in combination with a factor ambiguity representation in terms of the area of feasible solutions (AFS, [7, 19, 25]). The AFS theory deals with geometric objects as inner and outer polygons [15, 3, 17] which implement the constraints for the existence of nonnegative pure component factorizations.

For spectral data with a low noise level or for model data, those spectra/frequency channels (as represented by the rows/columns of the spectral data matrix) are considered to be essential which are related to the vertices of the inner polygons (polyhedra). The essential vertices of the inner polyhedra are related to the essential spectra and the essential frequency channels, respectively. The remaining data representing points of the inner polyhedron do not have an impact on the factor ambiguity. In this way, a reduction of the spectral data matrix to the essential rows/columns preserves the information content of the original matrix with respect to the pure components. The key role of these vertices can easily be understood as follows: The relative positions of the vertices of the inner polyhedron being embedded in the outer polyhedron determine the potential factor ambiguity within the geometric Borgen plot construction. In other words, the shape of these polyhedra restricts the possible Borgen triangles/simplices [10, 17, 20].

Unfortunately, the inner polyhedra are well-known to be sensitive to noise. For noisy data some of the vertices of these polyhedra can be scattered across the abstract space, even distant to their positions in the noise-free case. This finding can easily be verified by studying model problems with and without noise. Sometimes, in the case of (small) negative data entries, the inner polyhedron can even intersect the boundary of the outer polyhedron; we present an example in Section 3. Then an AFS construction is impossible. The remedy is as follows: In a first step, the outer polyhedron

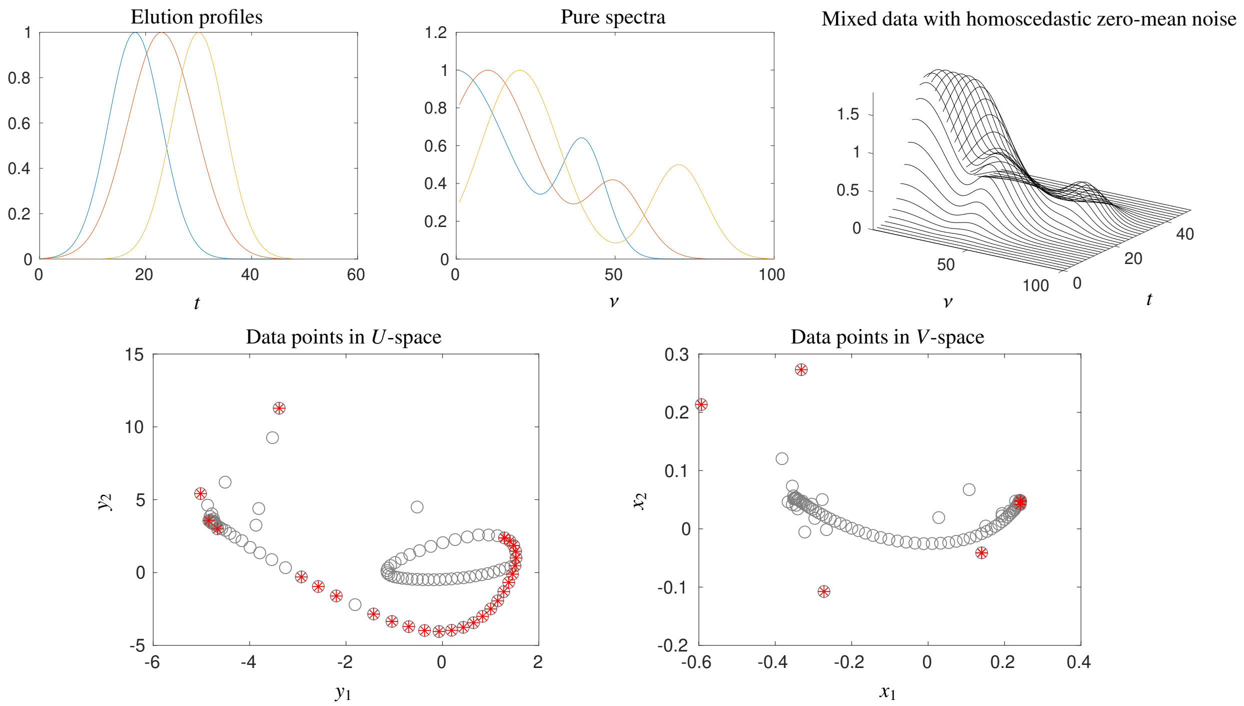

Figure 1: A chromatographic model problem illustrating data essentiality by the vertices of the inner polygon. Homoscedastic, zero-mean and low-level noise with standard deviation 10^{-3} has been added. The two upper plots show the elution profiles, the pure component spectra and the mixed data. The two lower plots show the associated low-dimensional data representations in the U- and V-spaces. The vertices of the inner polygons are marked by red stars.

is constructed, and this construction is known to be stable even for noisy data and also for data with small negative entries. In a second step, the inner polyhedron is constructed by means of duality arguments on the basis of the already known outer polyhedron. This construction results in a more reliable inner polygon construction. In the noise-free case the new approach coincides with vertex detection. The core of this paper is to compute essential information avoiding a direct evaluation of the inner polyhedron by using the outer polyhedron with its improved stability and duality. This work suggests a fast computational approach for finding the essential parts of spectral data sets which avoids multiple and costly computations of the outer polyhedron as used in our former work [21]. Moreover, the new approach is applicable to chemical systems with an arbitrary number of chemical species.

1.1. Existing approaches for noisy data and their limitations

The following strategies have been investigated for the detection of essential spectral information.

1. The approach [21] finds so-called *relevant* spectral information by multiple outer polygon computations by using inverse polygon inflation. The suggested algorithm is stable for noisy data even in the presence of small negative entries. However, the method is restricted to data sets including not more than three chemical species. Its computing time grows quadratically in the dimension of the input data due to the high number of internal outer polygon computations and subsequent comparison operations.

2. The aforementioned restriction to systems with three species at most can be bypassed if the approach [21] is applied to a low-rank approximation of the data set spanned by the three dominant singular values and the associated singular vectors. Such a projection technique has the disadvantage that potentially important information of the non-reduced spectral mixture data is not taken into account.

3. The approach presented in [5, 18] suggests to continue the usage of the vertices of the inner polyhedra for indicating the essential spectra also for noisy data. If all entries of the spectral mixture data matrix are far away from zero, then this detection of essential information is stable also in the presence of noise. However, absorption values close to zero (after subtraction of background signals) may result in a wrong detection of the essential information, since then the inner polyhedron is susceptible to interference.

These three techniques to identify essential spectral information have specific advantages, disadvantages and limitations. Next we apply these techniques to two data sets and critically discuss the results.

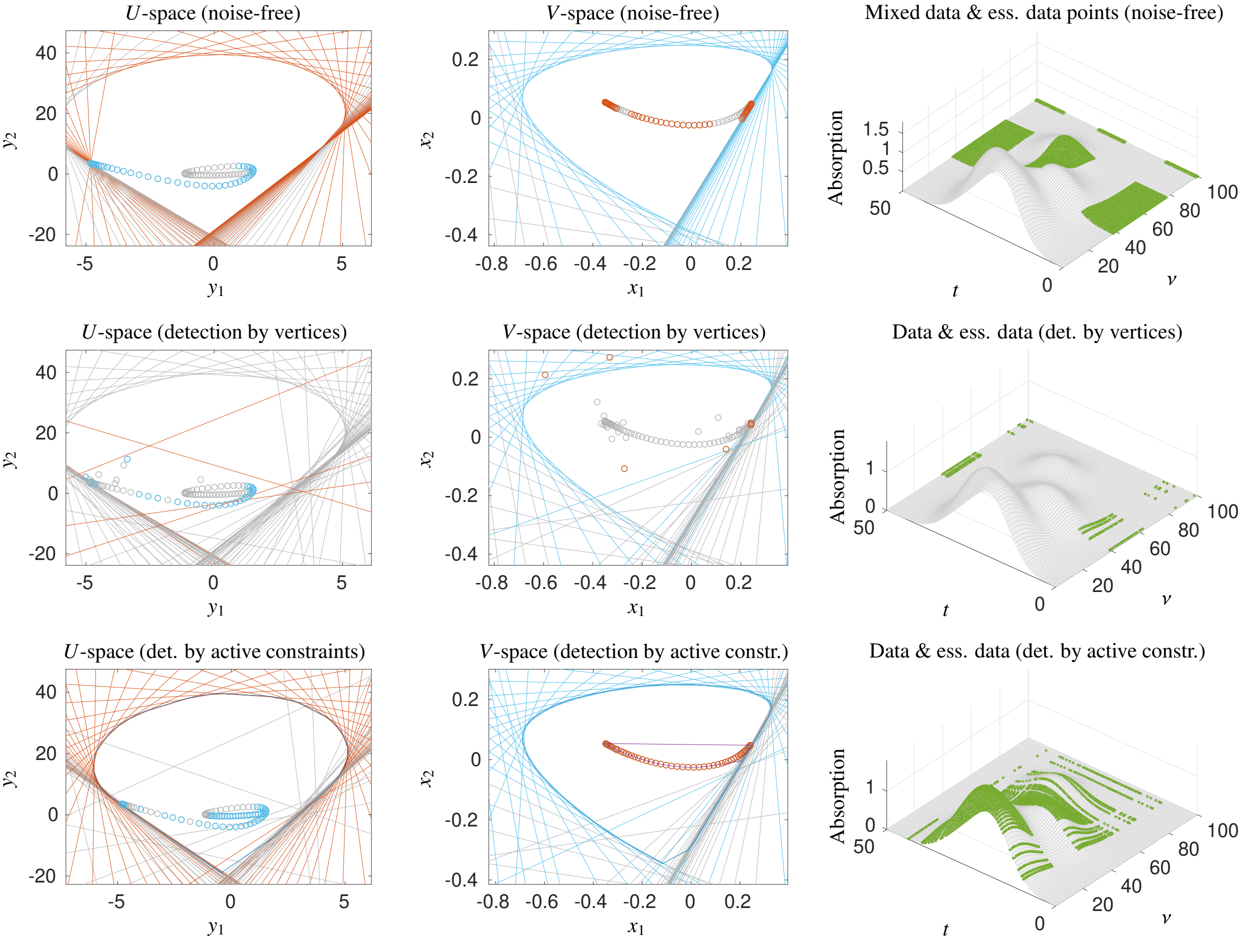

Figure 2: Comparison of essential data analyses for the chromatographic data set as introduced in Sec. 1.1. Essential parts of the spectral data are marked green within the absorption data grid (gray). Top row: Results for noise-free data. Some, but not all regions of dominant absorption are marked green. These regions are considered as essential in the noise-free case. 2nd row: Results by approach 3 for noisy data. This approach uses the vertices of the inner polygons for the detection of essential data. The results suffer from small negative entries; none of the major peaks is marked as essential. Bottom: The new approach finds active nonnegativity constraints in a first step. These active constraints are marked by colored lines (blue and red) in the two AFS plots. In a second step, duality allows to conclude from active constraints to the associated dual data points which are marked by blue and red circles. The green areas in the data plot (bottom, right) indicate that the areas of dominant absorption are determined as essential. Left and centered columns: For all analyses the detected essential data points and the dual active nonnegativity constraints are marked in red and blue. If data does not correlate to active nonnegativity constraints, then it is marked by gray lines and circles in the U- and V-space.

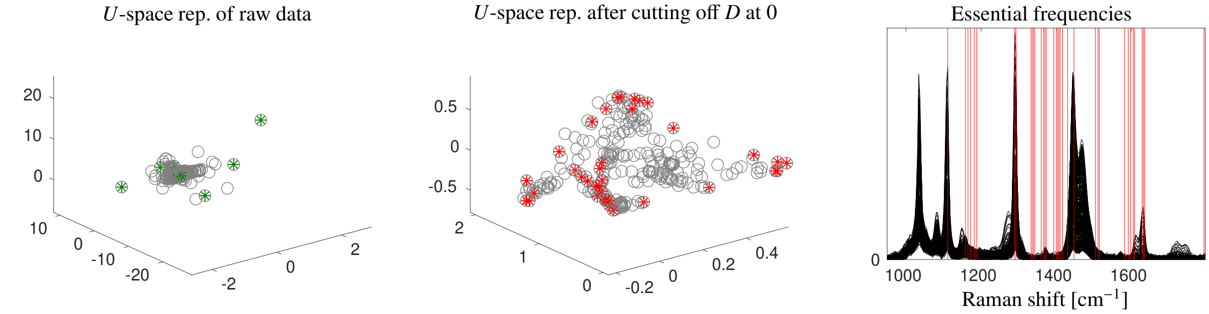

Figure 3: An experimental, noisy Raman data set of an oil-in-water emulsion with four ($s = 4$) chemical species. Left: The data representatives in U-space are marked by gray circles for the measured data after background subtraction. The vertices of the 3D inner polyhedron are marked by green stars. Small negative matrix elements are responsible for relatively large expansion coefficients, see the axes values. Center: Data representatives in the U-space after truncation of all negative entries in D. Then all expansion coefficients are much smaller (as all data is representable within the outer polyhedron). Again, the vertices of the inner polyhedron are plotted by red stars. Right: The series of spectra (black, only every 10th spectrum plotted). All frequency channels that are associated with vertices (red stars) of the inner polyhedron are marked by vertical red lines. Some frequency channels with dominant signals seem to be detected in a correct way, whereas other dominant peaks do not belong to essential frequency channels.

1. First, we consider a chromatographic model problem with $s = 3$ chemical components. Its elution profiles are modeled as shifted Gaussians. We add normal distributed, homoscedastic zero-mean noise with the standard deviation 10^{-3}. Thus the noise level is rather low. After noise addition all negative entries of D are truncated to zero. Fig. 1 shows the elution and spectral profiles and their low-dimensional representations in the U- and V-space. The vertices of the inner polygons are marked by red stars. We observe that the essential information detection in V-space is not stable since the essential information for noisy data is very different from the essential parts for the noise-free data. The reason is that due to the form of the elution profiles for several of the measured spectra all absorption values are close to zero. For close-zero rows of D the impact of homoscedastic noise is relatively large and this does also hold for the data representatives in the U- and V-spaces. Fig. 2 illustrates for this data set the essential data analysis by approaches 1 and 3 for the noise-free case (top row of plots) and after noise addition. The second row of Fig. 2 shows the results of an application of approach 1 where the essential (green) regions do not relate to regions of major absorbance, but to regions with small, even negative data components. The criteria on essentiality sensitively respond to these small entries. This indicates a potential weakness of this approach. In contrast to this, approach 3 extracts for the same noisy data essential data regions which correlate much better to the major absorbing regions.
2. Second, we consider an experimental Raman hyperspectral image of an oil-in-water emulsion from [5] with $k = 3600$ spectra and $n = 253$ spectral channels. A detailed analysis of this data set is contained in Sec. 4.2. Negative entries in the data are truncated, otherwise the results are even worse. We assume $s = 4$ species for this data set. The associated data representing points which span an inner 3D polyhedron in the U-space are plotted in Fig. 3. Again, approach 3 determines the vertices of the inner polyhedra. The detected essential frequency channels are marked by red stars in the U-space and also by red lines in the plot of the Raman spectra. The results are inconsistent. Some frequency channels are correctly detected and others that seem to belong to the essential data are omitted. We do not state a one-to-one relation of dominant signals and essential data.

The found inconsistencies justify a revised study of the data essentiality criteria for noisy data.

1.2. Contents and organisation of the paper

This work presents a new approach for a stable and computationally fast detection of essential spectra and frequency channels in noisy spectral mixture data. The method is applicable to chemical systems with any number of chemical species. The key ingredient of this method is the identification of active nonnegativity constraints. These active nonnegativity constraints, which determine the shape of the outer polyhedron, are related by duality arguments to the essential vertices of the inner polyhedron. As vertices of the inner polyhedron determine essential spectral information, the analysis can be restricted to active nonnegativity constraints. Throughout the paper we compare the results for the new algorithm with the ones for approach 3 from Sec. 1.1. In some situations the results coincide and in others the results differ.

The paper is organized as follows: Sec. 2 reviews the theory of low-dimensional representations in the spaces of left and right singular vectors and explains the duality between inner and outer polyhedra. In Sec. 3 we discuss an approach of how to work with noise and small negative entries in MCR factorizations and introduce the new approach to identify essential information of a data set. Finally, Sec. 4 presents results for experimental IR data, spectroelectrochemical data and hyperspectral imaging Raman data. We also discuss approximation methods which work with low-rank approximations of the full data matrix by artificially setting s smaller than the number of anticipated chemical species.

2. A geometric approach to the essential spectral information

Let $D \in \mathbb{R}^{k \times n}$ be the spectral data matrix with k being the number of spectra (samples) and n the number of spectral channels. The number of anticipated, independent chemical species is denoted by s. Then the truncated SVD of D reads $U\Sigma V^{\mathrm{T}}$ where the s columns of the $k \times s$ matrix U and the $n \times s$ matrix V contain the left and right singular vectors of D which correspond to the s largest singular values of D. The multivariate curve resolution problem (MCR) aims at computing a chemically interpretable, approximate factorization $D \approx CS^{\mathrm{T}}$. Therein, C is the matrix of concentration profiles along the time axis, and S is the matrix of the associated pure component spectra. These factors can be represented in terms of an invertible $s \times s$ matrix T of expansion coefficients so that [13, 14]

$$C = U\Sigma T^{-1}, \qquad S^{\mathrm{T}} = TV^{\mathrm{T}}. \tag{1}$$

This factorization has an inherent factor ambiguity, which is also called a rotational ambiguity, see [28, 25] and others.

Essential spectral information is deeply connected with the factor ambiguity. Whenever spectral information reduces the factor ambiguity, then this information is considered to be essential [21, 5]. Ideally, the spectral information is so strong that only a single (aside from scaling) factorization exists. Continua of possible factorizations exist in most cases. This is the reason why we are seeking for criteria based on the structure of the spectral data which can reduce the factor ambiguity. Next we discuss certain geometric objects which determine the factor ambiguity.

2.1. Inner and outer polyhedra

Pairs of an inner and outer polyhedron each for the factors C and S are the key objects for determining the factor ambiguity of an MCR problem, see for instance [7, 19, 25]. For systems with s chemical species the polyhedra lie in an $(s-1)$-dimensional space. Next we discuss the polyhedra for the factor S. First, there is an outer polyhedron which reads in the noise-free case

$$\mathcal{F} = \left\{x \in \mathbb{R}^{s-1} : (1, x^{\mathrm{T}})V^{\mathrm{T}} \geq 0\right\}.$$

In words, the polyhedron represents all points x so that $a = (1, x^{\mathrm{T}})V^{\mathrm{T}}$ is a componentwise nonnegative vector. Any feasible spectral profile must be nonnegative, and hence, the associated x must be located in $\mathcal{F}$. A second inner polyhedron $\mathcal{I}$ is the convex hull of the data representing vectors

$$a_i = \frac{(DV)^{\mathrm{T}}(2:s,i))}{(DV)_{i1}} = \frac{(U\Sigma)^{\mathrm{T}}(2:s,i)}{(U\Sigma)_{i1}}, \quad i = 1,\ldots,k. \tag{2}$$

Thus the $a_i \in \mathbb{R}^{s-1}$ are the representatives of the measured spectra (namely the rows of D) in the V-space since a_i can also be written in the form

$$a_i = \frac{e_i^{\mathrm{T}} D[v_2,\ldots,v_s]}{e_i^{\mathrm{T}} D v_1}$$

where e_i is the ith standard basis vector in the k-dimensional space and $v_1,\ldots v_s$ are the right singular vectors. The outer polyhedron encloses the inner polyhedron in the sense of the set enclosure $\mathcal{I} \subset \mathcal{F}$. The inner and outer polyhedra for factor C are similarly defined or are accessible by applying the definitions above to the transposed spectral data matrix D^{T}. We refer to [20, 25] for their explicit definitions.

A pair of an inner and and outer polyhedron determines the factorization ambiguity of the associated factor. The key property is as follows: the s vertices of any simplex in the $(s-1)$-dimensional space which encloses the inner polyhedron and which is enclosed in the outer polyhedron determine the feasible profiles of the s chemical species. This leads to the definition of the AFS being the set of all vectors in the U- or V-space whose corresponding spectral or concentration profile can be extended to a complete nonnegative factorization $D = CS^{\mathrm{T}}$ with $C, S \geq 0$. As the factorization problem involves two factors, there are two AFS-sets, one for each factor. A further prominent global approach for describing the factor ambiguity consists of the computation of band boundaries by minimizing and maximizing the signal contribution function [4, 27].

2.2. Noise-free data and essential spectral information

For high-dimensional data, the question arises which measurements (entries of D) are important with respect to the extent of the factor ambiguity and which are not. Independently, in [5] and [21], the terms essential and relevant have been introduced in order to categorize certain spectra or frequency channels in this sense. Here we use the term essential. However, the new approach differs from [5] in a sense that it focuses on active constraints of the dual factor.

Definition 2.1. *For noise-free data a spectrum or a frequency channel and its representing point in the U- or V-space are called essential if its addition or removal has an impact on the shape of the respective outer polyhedron. Then this data point is a vertex of the inner polyhedron. Otherwise, the point is called non-essential.*

The identification of the essential parts of the spectral data is based on the duality between the inner and outer polyhedra [10, 17, 20, 22]. In particular, the vertices of one inner polyhedron are dual to the facets of the other outer polyhedron. Fig. 4 illustrates the duality relation of a vertex (namely the data point according to channel 39) of the inner polygon in the U-space to its dual line touching the outer polygon in the V-space. We use the first data set as introduced in Sec. 1.1 without noise. The figure also shows the feasible profiles attaining the value 0 at ν_{39} as minimal entry, which corresponds to an active nonnegativity constraint. Computation of profiles using active constraints for certain channels has also been proposed in [16].

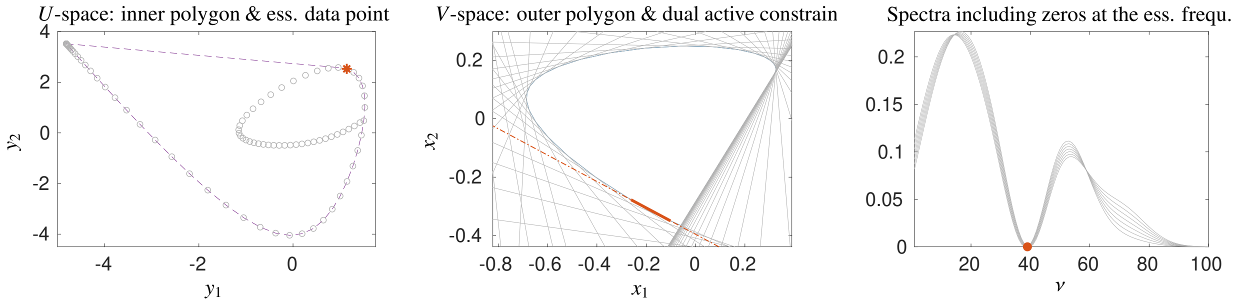

Figure 4: The idea of active constraints explained for noise-free model data from Sec. 1.1. Left: The data points in the U-space are marked by gray circles. Data point no. 39 is selected (red star, essential point). Center: The gray lines restrict the outer polygon in the V-space. The red dash-dotted line belongs to channel 39 and represents an active constraint. Its line segment touching the outer polygon is drawn bold and red. Right: Spectral profiles belonging to the bold red line segment. Each profile is zero at ν_{39} since the constraint is active, thus (5) is fulfilled.

For noise-free data it is simple to check the essentiality of a measured spectrum or a frequency channel. The first step is a computation of the inner polyhedron of the factor C by evaluating the convex hull of the vectors (2) representing the rows of D. See [20, 25] for the analog vectors in the case of the factor S where the columns of D are considered. In MatLab a convex hull computation is possible with the routine convhull. Finally, a spectrum $D(i, :)$ is essential if and only if the data representing vector a_i is a vertex of $\mathcal{I}$. Further, a frequency channel is essential if and only if its low-dimensional representation is a vertex of the inner polygon in the U-space [21, 5, 18].

2.3. Noisy data, essential spectral information and low-rank approximations

We subsume under the term data points the representing points in the U- or V-plane of the spectra (rows of D) or frequency channels (columns of D). For noisy data we consider very different processes to classify data points as essential or non-essential. In an analogous way the classifications in terms of relevance in [5] and [21] are the same for noise-free data and are different for noisy data. In [5] the vertices of the inner polygons are taken in a straightforward way as the essential data points. In contrast to this, [21] additionally applies a routine based on the relative consideration of errors as in the polygon inflation algorithm.

Definition 2.2. *For noisy data a spectrum or a frequency channel is called essential if its addition or removal has an impact on the generalized outer polygon whose definition relies on the weakened nonnegativity constraints* (3). *The associated data point is also called essential.*

Remark 2.3. *For noisy data it is difficult to speak of a "correct" or "wrong" classification of data points as the criteria which are introduced for noise-free data cannot directly be applied to noisy data. Depending on the classification method certain intermediate steps must be applied in order to form an approximate setup to which the classification criteria can be applied. Hence, referring to correct or wrong classification for noisy data should always be combined with a reference to the classification method.*

We have already shown that approaches 1 (due to the large number of $k + n + 2$ outer polygons which are to be computed) and 3 have some disadvantages. Next, Theorem 2.4 points out the projection-based approach 2 can also miss essential data points.

Theorem 2.4. *Let $\mathcal{F}$ be the outer and $\mathcal{I}$ the inner polyhedron in the $(s-1)$-dimensional space of a spectral data matrix D with the rank s. Further let $\mathcal{F}'$ and $\mathcal{I}'$ be the outer and inner polyhedron with respect to the rank-m approximation D' of D with $m < s$. (See Sec. 2.4.2 of [9] for SVD-based low-rank approximations.)*

Then it holds that:

1. *The set of coordinates of the outer polyhedron $\mathcal{F}'$ with respect to the low-rank approximation after coordinate-wise extension by $s - m$ zeros equals the intersection of the set of coordinates of the initial polyhedron $\mathcal{F}$ with the $(m-1)$-dimensional linear subspace $\mathcal{H} = \{x \in \mathbb{R}^{s-1} : x_m = \ldots = x_{s-1} = 0\}$.*
2. *The inner polyhedron $\mathcal{I}'$ with respect to the reduced basis (and after coordinate-wise extension by zeros) equals the projection of $\mathcal{I}$ on $\mathcal{H}$.*

Proof. For 1.: The outer polyhedron with respect to the reduced basis spanned by m singular vectors reads

$$\mathcal{F}' = \{z \in \mathbb{R}^{m-1} : (1, z^{\mathrm{T}})(V(:, 1 : m))^{\mathrm{T}} \geq 0\}.$$

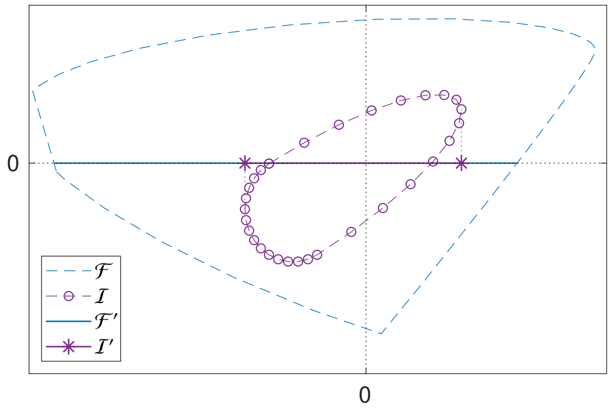

Figure 5: Graphical illustration of the items 1 and 2 of Thm. 2.4 for the case of three species so that $s = 3$. Then outer and the inner polygon are two-dimensional. If only two left and right singular vectors belonging to the two dominant singular values are used, then the outer polygon $\mathcal{F}'$ is the intersection of $\mathcal{F}$ with $\mathcal{H} = \{x \in \mathbb{R}^2 : x_2 = 0\}$ (the x_2-axis) namely the bold, blue, horizontal line. In contrast to this construction by intersection, the inner polygon $\mathcal{I}'$ is the projection of $\mathcal{I}$ to $\mathcal{H}$ reduced to its first coordinate.

Thus $\mathcal{F}'$ is the intersection (and not the projection) of the two sets of the outer polyhedron $\mathcal{F}$ and $\mathcal{H}$. Fig. 5 illustrates this.

For 2.: The inner polygon $\mathcal{I}'$ is the convex hull of the vectors

$$a'_i = \frac{(DV)^{\mathrm{T}}(2:m,i))}{(DV)_{i1}}, \qquad i = 1, \ldots, k.$$

Therein the vector a'_i is the representation of the ith row of D with respect to the first m singular vectors. Thus a'_i is the projection of a_i to the linear subspace $\mathcal{H}$. Consequently, the convex hull $\mathcal{I}'$ of all vectors a'_i is the projection of $\mathcal{I}$ on $\mathcal{H}$, see Fig. 5. □

3. Stable detection of essential data points for noisy data by soft constraints and duality

For noisy data the direct convex hull computation of the inner polyhedron is known to be unstable so that a subsequent AFS computation in the U- and V-space may become impossible. Fig. 6 shows a typical example for which the inner polygon intersects the boundary of the outer polygon. Our strategy is to avoid a direct computation of the inner and outer polygons (or polyhedra) by hard constraints. Instead, the outer polygons are constructed by using soft nonnegativity constraints, which allow small negative entries. Subsequently the inner polyhedra are computed by means of duality-based techniques as dual objects to the soft-constrained outer polyhedra. Algorithms as inverse polygon inflation for $s = 3$ and ray casting for any $s \geq 2$ work very well in order to approximate the outer polyhedra in the presence of noise [23, 24]. The idea is to detect which frequency channels/spectra are associated with active constraints in the outer polyhedron construction and which belong to inactive constraints. Duality theory predicts that active constraints of one outer polyhedron correspond to the vertices of the dual inner polyhedron [22, 21]. The suggested approach is different to [21], much faster and applicable to chemical systems with any number of chemical species.

3.1. MCR factorizations in the presence of noise

The MCR problem for noise-free data deals with the computation of the strictly nonnegative factors C and S so that $D = CS^{\mathrm{T}}$. For noisy data various approaches are available to compute reliable approximate pure component factorizations. One approach is to compute strictly nonnegative factors C_+ and S_+ but to accept approximate factorizations of the form $\left\|D - C_+S_+^{\mathrm{T}}\right\| < \varepsilon$ where ε is a small positive control parameter. Computational implementations of this approximate-factorization approach can use the nonnegative least squares algorithm in order to calculate approximate factors C_+ and S_+. Alternatively, we can use the SVD factorization by Eq. (1) which is combined with a subsequent truncation of all negative matrix elements [8]. For noisy data or after background subtraction it is more appropriate

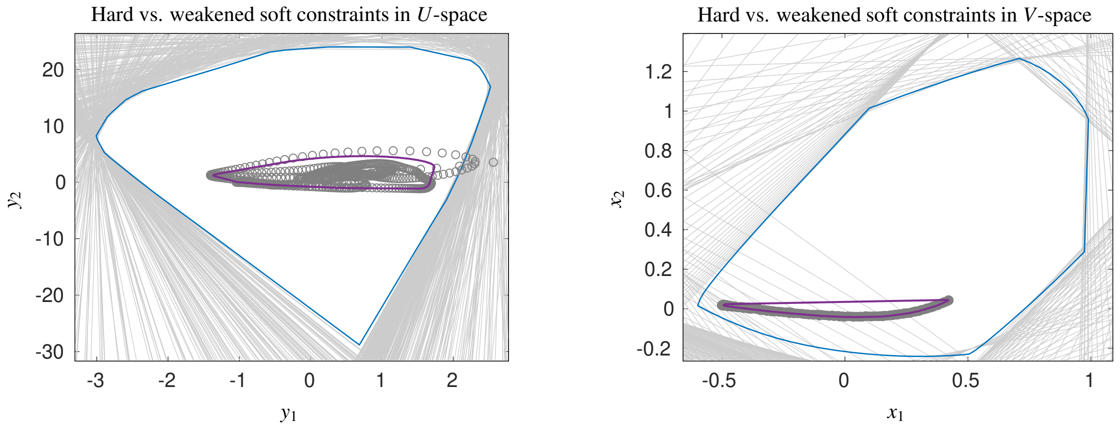

Figure 6: Comparison of the inner (purple) and outer (blue) polygons for hard and soft constraints in application to the noisy, experimental IR data set from [12]. The gray lines represent the hard nonnegativity constraints, and the gray circles represent the data points (columns and rows of the spectral data matrix). These hard constraints do not allow even a single solution since the data points leave the outer polygon (left plot) and intersect the hard nonnegativity constraints (right plot). Additionally, the plots show the soft-constrained approximations to the outer and the inner polygons (in blue and purple) as introduced in (3). For these approximations feasible triangles exist which are located in the outer polygon and which include the inner polygon. Therefore, feasible pure component factorizations can be constructed.

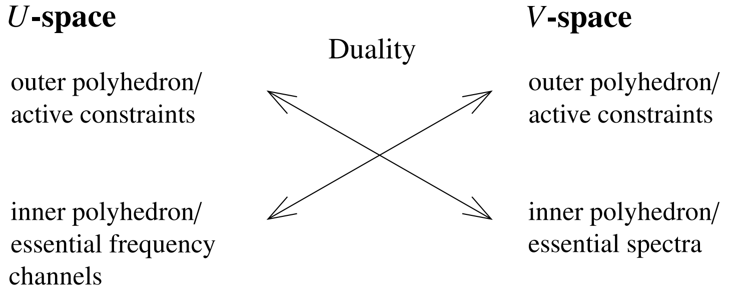

Figure 7: Duality relations between inner and outer polyhedra in the U- and V-space with respect to active constraints and essential spectral information.

to accept small negative matrix elements of C and S whose size can be bounded in terms of the following relative measures

$$\frac{\min C(:,i)}{\max |C(:,i)|} \geq -\varepsilon, \qquad \frac{\min S(:,i)}{\max |S(:,i)|} \geq -\varepsilon \tag{3}$$

for $i = 1, \ldots, s$. Again, ε is a small positive control parameter. These weakened nonnegativity constraints can be combined with other penalty functions and can jointly be enforced by a numerical optimization process. In words, a factorization is accepted if for each profile the absolute value of the smallest (negative) entry is not greater than ε times the maximum of the entire profile. The polygon inflation algorithm [23], the ray casting method [24] and the dual Borgen plot construction [20, 22] work with this approach. In all these cases the factors C and S are represented as in Eq. (1). Fig. 6 illustrates how these weakened nonnegativity constraints work. With hard nonnegativity constraints, the inner polygon leaves the outer polygon in the left plot, and some hard nonnegativity constraints intersect the inner polygon (right plot). Feasible factors do not exist. In contrast to this, soft nonnegativity constraints together with a duality-based inner polygon construction allow ranges of feasible solutions.

3.2. Detection of active constraints

Next we pursue the goal of identifying those spectra and frequencies that represent active constraints of the outer polyhedra. We focus on the outer polyhedron in the V-space and compute essential frequency channels. The process can easily be adapted to the outer polyhedron in the U-space for detecting essential spectra. See Fig. 7 for a graphical representation of these relations.

Let $\mathcal{F}$ be the noisy data approximation of the outer polyhedron of the factor S according to (3). Further let $F \in \mathbb{R}^{N\times(s-1)}$ be the matrix which comprises an N-point sampling of $\mathcal{F}$ containing row-wise a list of N sampling boundary points.

Then the jth row $x^{(j)} = (F(j,:))^{\mathrm{T}}$ of F represents the associated spectral profile

$$a^{(j)} = \left(1, (x^{(j)})^{\mathrm{T}}\right) V^{\mathrm{T}}. \tag{4}$$

As $x^{(j)}$ is an element of the outer polyhedron, namely $x^{(j)} \in \mathcal{F}_S$, it holds componentwise that $a^{(j)} / \max |a^{(j)}| \geq -\varepsilon$. Since $x^{(j)}$ is located on the surface of $\mathcal{F}$, equality is attained at least in one component of the last inequality so that

$$\frac{\min_{\ell}(a_{\ell}^{(j)})}{\max |a^{(j)}|} = -\varepsilon. \tag{5}$$

Typically, $x^{(j)}$ is only an approximation to a boundary point. Further, (5) holds only in an approximate manner with respect to the precision as used for the computation of $x^{(j)}$. If (5) is attained in an index ℓ, then ℓ belongs to a facet of $\mathcal{F}$. Thus the associated restriction is active and the frequency channel ℓ is essential. For a certain $x^{(j)}$ equality (5) can be fulfilled for more than a single index within the given precision of the approximation. All such active constraints are collected in a set, see the next subsection. If the surface has been scanned sufficiently well, then a subset of all sampled points $x^{(j)}$, $j = 1, \ldots, N$, is the desired list of essential indices and frequency channels.

3.3. *Numerical precision of the boundary approximation*

The surface sampling (by inverse polygon inflation as in FACPACK [24] or by ray casting if more than three chemical species are expected) results in a discrete mesh of points $x^{(j)}$, which are all numerical approximations of the precise surface, since the iterative optimization includes some termination criteria. The optimization is built around the simple and very stable bisection method. Due to the finite precision termination criteria we cannot expect that numerically computed points $x^{(j)}$ fulfill (5) exactly or to be located exactly on the surface of the polyhedra. The suggested numerically stable algorithm classifies a certain index $\ell \in \{1, \ldots, n\}$ to be essential if

$$\frac{a_{\ell}^{(j)}}{\max |a^{(j)}|} \leq -\varepsilon + \delta_{\text{bnd}} \tag{6}$$

for a given proper control parameter $\delta_{\text{bnd}} > 0$. We suggest to use $\delta_{\text{bnd}} = 1.01 \left|\frac{\min(a^{(j)})}{\max(a^{(j)})}\right|$. Finally, for each boundary point index $j \in \{1, \ldots, N\}$ we compute the set of associated active constraint indexes

$$I^{(j)} = \{\ell : \text{ inequality (6) is fulfilled for } \ell\}. \tag{7}$$

This is the set of essential frequency channels and is, by definition, a subset of the set $\{1, \ldots, n\}$ of all frequency channel indexes.

3.4. *Index in which a profile takes its maximum is also essential*

The evaluation criterion (6) works with relative measures based on normalized profiles. This normalization is a necessary intermediate step to work with noisy data. We have observed that the index in which the profile $a^{(j)}$ takes its componentwise absolute maximum $\max |a^{(j)}|$ can also belong to an essential spectrum or frequency channel. Skipping such an index ℓ with $\ell = \arg\max |a^{(j)}|$ would move $x^{(j)}$ to the outside of $\mathcal{F}$ and then ℓ would gain essentiality. Hence we define a final set of essential indexes as the union of all sets $I^{(j)}$ by (7) together with the indexes in which the maxima are taken and written as $\arg\max \left|a^{(j)}\right|$. So we get

$$M = \bigcup_{j=1}^{N} \left(I^{(j)} \cup \arg\max \left|a^{(j)}\right| \right).$$

3.5. *Application to the chromatographic model problem*

We determine the essential information for the chromatographic model problem as presented in the introduction, see also Fig. 1. The two left plots of Fig. 8 show the U- and V-space representations of the detected essential data points for the threshold value $\varepsilon = 10^{-3}$. In contrast to a straightforward usage of the vertices of the inner polygons the duality-based outer polyhedron strategy does not select data points corresponding to noisy spectra/frequency channels close to the zero vector; compare with Fig. 1 and Fig. 8. However, the set of detected essential data points could be smaller. In other words, several data points are classified as essential although they are clearly not close to any vertices

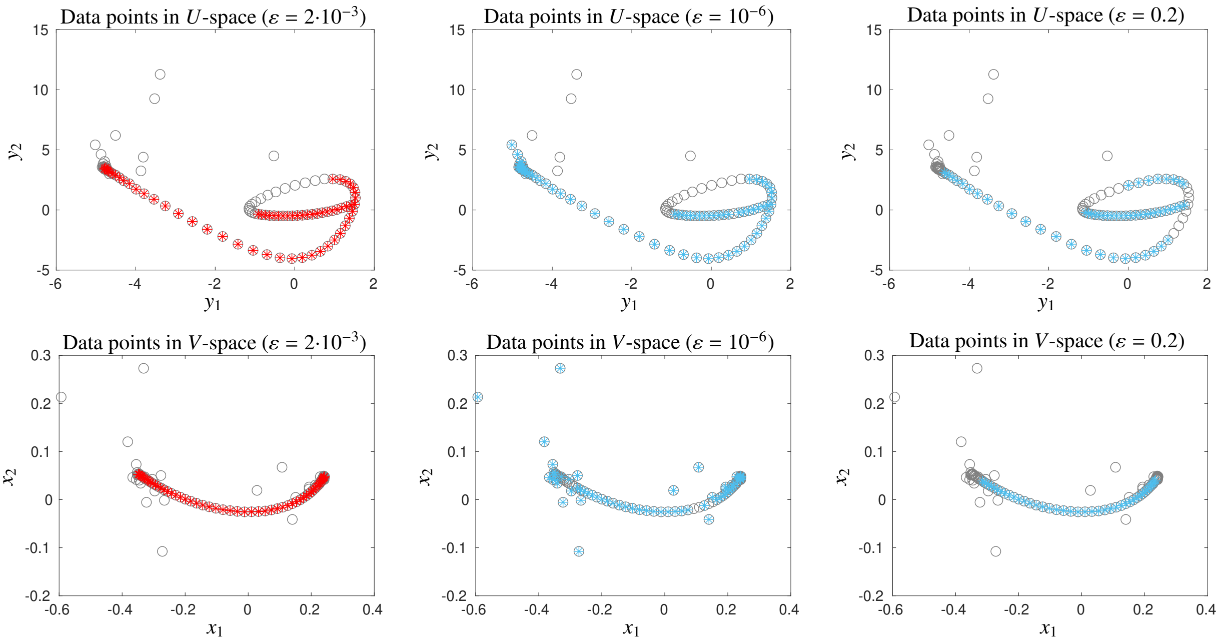

Figure 8: Two left plots: Essential data points (marked by red stars) in the U- and the V-space for the model problem as introduced in the introduction and shown in Fig. 1. The essential data points are detected by the active constraints strategy based on the dual outer polygons and for the threshold value $\varepsilon = 2{\cdot}10^{-3}$. The essential data points are located on smooth curves, some of them seem to have no impact on the inner polygon. In contrast to Fig. 1 no outliers indicate poor or even wrong classifications. Center and right: The inappropriate settings $\varepsilon = 10^{-6}$ (too small) and $\varepsilon = 0.2$ (too large) result in data point classifications (blue) being either too sensitive (two centered plots) or too coarse (two right plots).

of the inner polygons. We consider this as a compromise made in order to avoid wrong classifications or a loss of information. The centered and right plots in Fig. 1 show how a too small or a too large setting of the threshold value ε can affect the results, namely with a sensitivity that is too large or too small.

We use the control parameter $\varepsilon = 2.5 \cdot 10^{-3}$ in response to homoscedastic and normal distributed noise with the mean value 0 and a standard deviation of 10^{-3}. Thus, statistically 99.4% of the noise values are larger than $-2.5 \cdot 10^{-3}$ and 98.76% of the absolute noise values are smaller than $\varepsilon = 2.5 \cdot 10^{-3}$. The setting $\varepsilon = 10^{-3} = \sigma$ appears to be too small since statistically only 15.87% of the noise-values would be smaller than ε.

3.6. *Computational efficiency and numerical stability*

The advantages of the suggested classification procedure are its efficiency and stability. The algorithm works with discrete approximations of the outer polyhedra in terms of surface samplings. We expect that the computational costs for increasing the surface resolution (number of points on the surface of the polyhedra) and for increasing the precision of the boundary approximation (distance to the true boundary) increases only linearly in the number of chemical species s.

4. Numerical studies

We apply the procedure for the identification of essential frequency channels and spectra to various experimental IR and UV-Vis data sets and also process and imaging data. A suitable setting of the number s of singular vectors which are considered for the computation (of the $(s-1)$-dimensional polyhedra) is crucial for a reliable identification of the essential spectral information. Next we test different parameter settings of s and of the control parameter ε according to (3) and compare the results concerning the identified essential frequency channels and spectra.

4.1. *Chemical image data for a three-component system*

The FTIR hyperspectral imaging data set was obtained by analyzing a three-component ($s = 3$) pharmaceutical powder composed of acetic acid, acetylsalicylic acid and caffeine, see [11, 5] for more details on the data. The pixel size was set at 25 x 25 μm and 32 scans were accumulated per spectral pixel. Only the fingerprint frequency range $675 - 1800\,\text{cm}^{-1}$ for the spectral measurements was selected for data analysis. The data is stored in an $M \in \mathbb{R}^{36\times36\times583}$

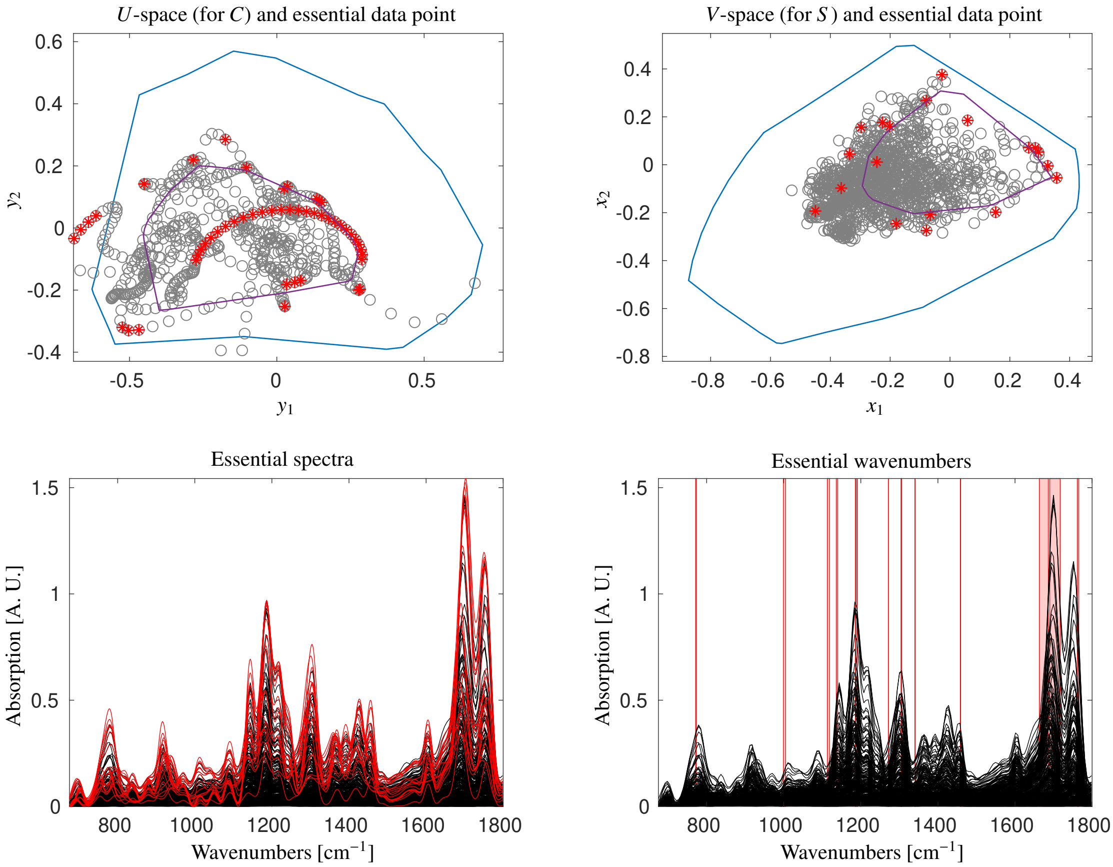

Figure 9: Analysis of experimental FTIR hyperspectral imaging data as introduced in Sec. 4.1. Upper plots: The data points are plotted in gray (all) and red (essential ones) in the U- and V-plane. The approximations of the outer polygons assuming noisy data conditions are drawn in blue and their associated dual inner polygons are drawn in purple. The control parameters are $\varepsilon_C = 0.035$ and $\varepsilon_S = 0.05$ and the precision for the boundary approximation in the polygon inflation algorithm is $\varepsilon_b = 10^{-6}$. The inner polygons are true subsets of the convex hulls of the data points since the control parameters shift the facets into the direction of the origin. However, many data points are outside the inner polygons. Lower plots: All spectra are drawn in black. The left plot shows the essential spectra ($k^* = 19$ out of $k = 1296$) in red. The right plot marks the essential wavenumbers ($n^* = 51$ out of $n = 583$) by vertical red lines.

field with $M(i, j, :)$ being the measurements at the coordinates $(\overline{x}_i, \overline{y}_j)$. Hence, the data sets includes $k = 36 \times 36 = 1296$ spectra.

The algorithm for identifying essential spectra/frequency channels uses the parameters $\varepsilon_C = 0.035$ for factor C and $\varepsilon_S = 0.05$ for factor S, see Eqns. (3) and (6). Fig. 9 shows the results. The detected essential data points are drawn in red, and the approximations of the outer (by inverse polygon inflation) and inner (by duality) polygons are drawn in blue/purple. Obviously, the essential spectra are neither a subset nor a superset of the convex hull of the vertices of the data points in the U- and V-space. The results appear to be consistent. In Fig. 9 the outlying essential data points are close to the boundaries of the respective polygons.

The impact of the control parameter settings for ε_C and ε_S on the outer polyhedra is illustrated in Fig. 10. For this purpose we mark the essential spectra indexes versus varying ε_C and also mark the essential frequency channels versus varying ε_S. The rule of thumb is as follows: With increasing parameter values, which allows for larger portions of negative entries, more parts of the spectral data are considered to be essential.

4.2. *Chemical image data with more than three components*

We consider Raman hyperspectral image data of an oil-in-water emulsion, see [5] for more details on this data. The sample consists of an emulsion base whose full chemical composition is complex and was originally described in [1]. Chemical interpretation and estimation of the number of components is further complicated by chemical interactions and physical changes occurring during the analysis of the mixture. The analyzed image consists of 60×60 pixels, so $k = 3600$, and $n = 253$ spectral channels in the range between 950 and 1800 cm^{-1}. Fig. 11 shows the data, its first 30 singular values and the first left singular vector. For the following analysis the control parameters $\varepsilon_C = 0.075$ and $\varepsilon_S = 0.025$ are used to bound negative entries in C and S according to (3). Due to noise around 21.2% of the entries

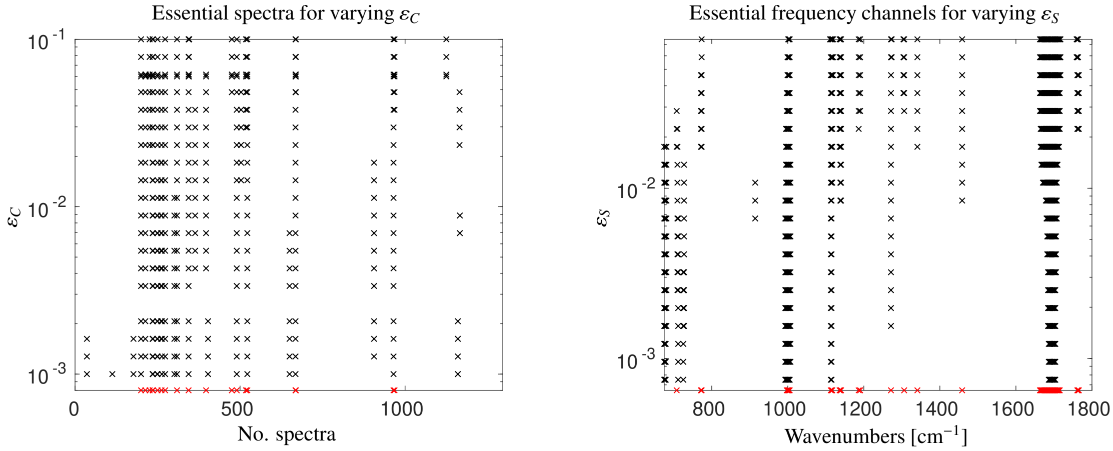

Figure 10: Dependence of the essential spectra/frequency channels on the control parameters ε_C and ε_S for the experimental FTIR hyperspectral imaging data as introduced in Sec. 4.1. Left: For varying ε_C the plot shows cross markers for the indexes of the essential measured spectra. Right: Indexes of essential frequency channels under variation of ε_S. The red markers on the bottom belong to the classification for $\varepsilon_S = 0.05$ and $\varepsilon_C = 0.035$. The given axis orientation allows easy comparability with the essential wavenumbers as plotted in Fig. 9.

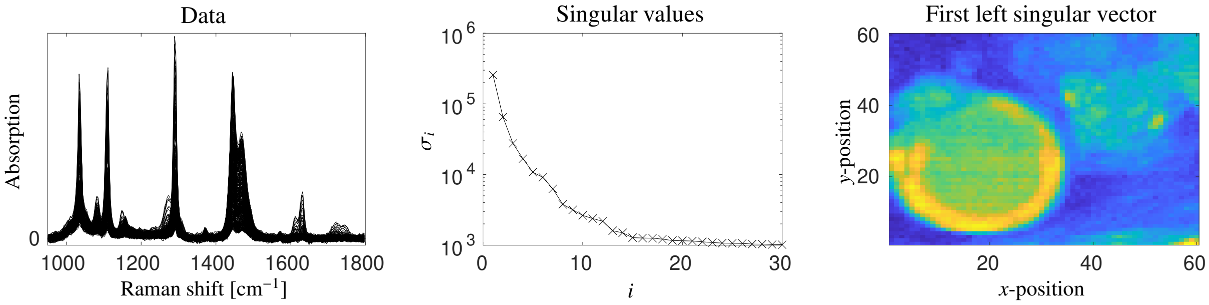

Figure 11: Raman hyperspectral image data as described in Sec. 4.2. The series of spectra (left), the first 30 singular values (center) and the first left singular vector (right). The curve of singular values does not clearly indicate the number of absorbing species (center plot) and the concentration values are not evenly distributed (right plot).

in the data matrix are negative after background subtraction. We do not remove negative entries from the data for the application of the essential data point analysis (in contrast to the procedure generating the plots in Fig. 3).

This data set suffers from spectral regions with minor absorption. Fig. 3 shows the low-dimensional U-space representation of the data together with the vertices of the inner polyhedron and their associated frequency channels. These results represent a partially doubtful and incomplete classification of the essential frequency channels (vertical red lines in Fig. 3). Several major peaks are not classified as essential. In contrast to this, the new algorithm based on the approximation of the outer polyhedron by means of soft nonnegativity constraints in combination with an active constraint detection and duality works in a better way. A number of $n^* = 18$ frequencies is detected as essential which includes all major peaks, see Fig. 12.

4.2.1. Projection to subsystems with fewer species

Option 2 in the introductory section suggests applying a projection approach for determining essential spectral information. For this purpose we consider the low-rank approximation of the spectral data matrix with the rank $s = 3$ or $s = 4$ even if one knows that the system involves more species. We test this strategy for the given Raman hyperspectral image data set of the oil-in-water emulsion. We cannot determine the true number of chemical species, take $s = 4$ intermediately as the number of species for this data set and analyze the rank-3 SVD approximation of the data set. Fig. 13 shows the data points in the 3D V-space of the original data by gray circles. This representation relates to the usage of four dominant singular values/vectors. For the rank-3 approximation of this data set we mark all essential data points with respect to $s = 3$ in the left subplot of Fig. 13 by red stars. The right plot shows all essential data points with respect to $s = 4$ by red stars. Obviously, many of these points are not located in the bottom plane; they have a clearly nonzero third component. This reveals a significant disadvantage of working with a number s that is too small since many essential data points are not detected. By a projection to the x_1-x_2-plane any information from the third coordinate x_3 gets lost (left plot). We state that the extracted essential information using $s = 4$ (right plot) covers

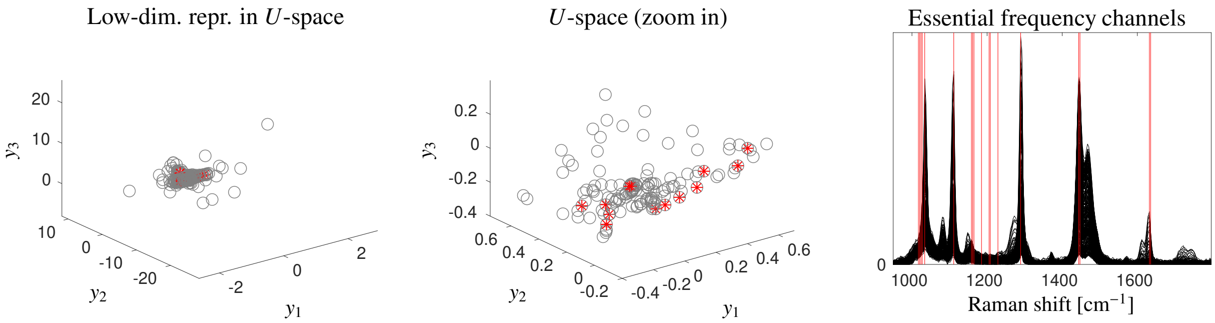

Figure 12: Detection of the essential frequencies for the Raman hyperspectral image data introduced in Sec. 4.2. The duality-based outer polyhedron approach works with the $s = 4$ singular vectors belonging to the four largest singular values. Left and center: data points in U-space (gray circles) and identified essential data points/essential frequency channels (red). The measured data has been preprocessed by a background subtraction, but without truncating negative entries of D. Right: The series of spectra (drawn in black). The found essential frequency channels are marked by vertical red lines. This result should be compared with Fig. 3 which shows the results of the approach based on the vertices of the inner polyhedron. The latter approach is related to the hard constraints $C, S \geq 0$.

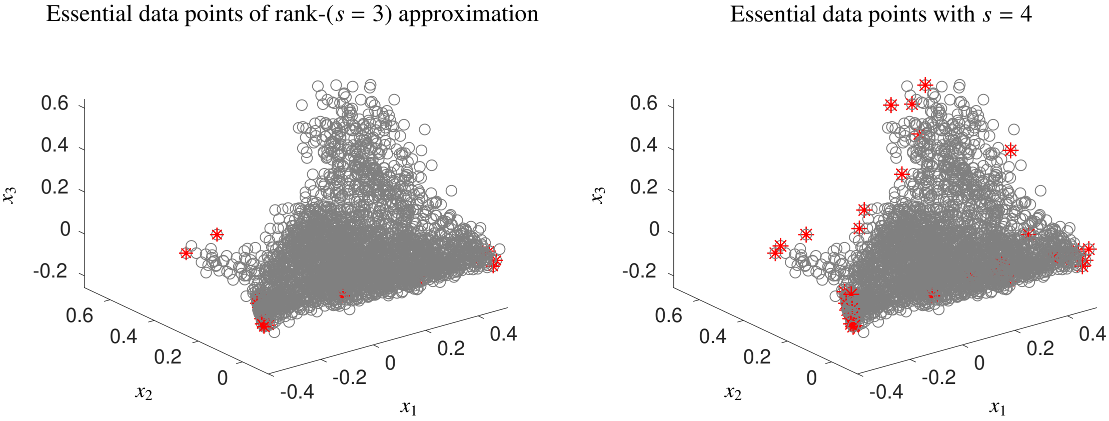

Figure 13: These two plots show the data representing points in the V-space by gray circles for the Raman hyperspectral image data introduced in Sec. 4.2. In the left plot we mark by red stars the essential data points which are found by a rank-3 approximation of the spectral data matrix again. The right plot shows the essential data points of the original data. Obviously, the projection of the data to $s = 3$ is far from being capable to capture all essential spectra.

all essential data points (inner polygon) and that these results are stable with respect to perturbations. If we denote by $\text{ess}_{C,s}$ the set of essential frequency channels computed with a rank-s approximation of the spectral data set and $\text{ess}_{S,s}$ the analog set of essential spectra, then we observe for the given data the (approximate) set enclosure relations

$$\text{ess}_{C,3} \subset \text{ess}_{C,4} \quad \text{and} \quad \text{ess}_{S,3} \subset \text{ess}_{S,4}$$

except for one element each related to the factors S and C.

4.2.2. *Stepwise increasing the number of assumed species and analysis of essential information*

Next we increase s and re-compute the essential data points. The plot of the singular values for this data set in Fig. 11 does not clearly indicate the number of absorbing species. Checking the left singular vectors (as this is done for $U(:,1)$ in the right plot of Fig. 11) suggests that only the first seven of them appear to have a non-oscillatory character and therefore potentially contain chemically relevant information. We compute the essential spectral information for $s = 3$ up to $s = 6$. See Fig. 14 for the results. The two left plots show the essential spectra indexes and the essential spectral channels for the different settings of s marked by crosses. By stepwise increasing s we observe that the set of essential indexes is more or less supplemented by new additional indexes, see items 2 and 3 of Thm. 2.4. This indicates some degree of stability of this approach, but also shows the incompleteness of the sets of essential indexes if s is too small. The right plot of Fig. 14 shows the positions of these essential data points with respect to the x-y image representation. See also [18] for a comparison of the present results with the approach based on the vertices of inner polyhedron and with $\varepsilon_C = \varepsilon_S = 0$, which means working with the hard constraints $C, S \geq 0$.

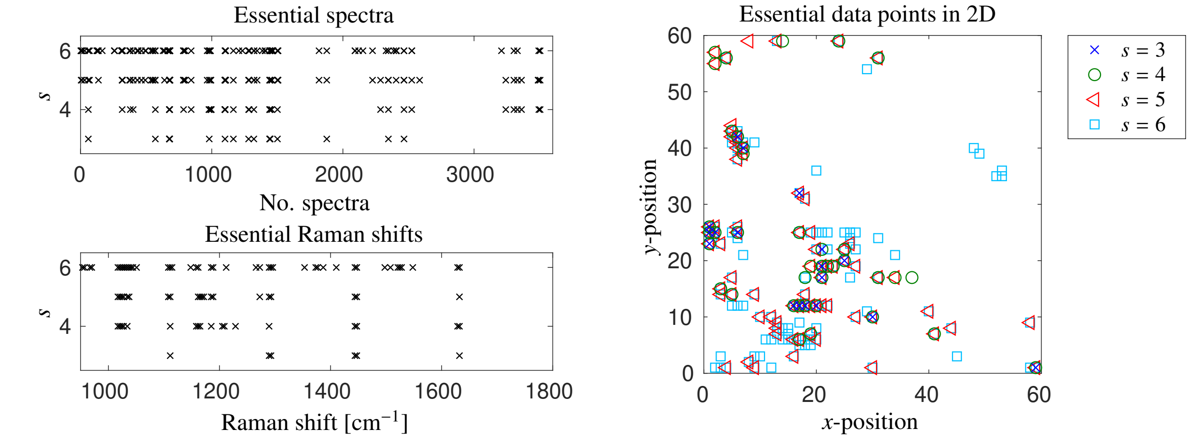

Figure 14: Essential spectra and Raman shifts (essential channels) for different values s for the Raman hyperspectral image data introduced in Sec. 4.2. Left: The essential spectra and essential Raman shifts are marked versus the spectra index respectively versus the Raman shifts for $s = 3, 4, 5, 6$. Right: The x-y-positions of the essential spectra are plotted for different values of s; compare this with the rightmost plot of Fig. 11.

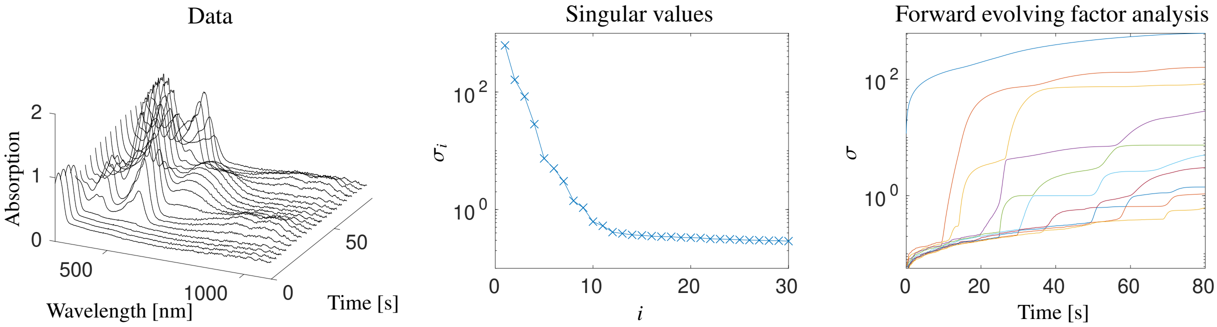

Figure 15: Spectroelectrochemical analysis of the anthraquinone system as introduced in Sec. 4.3. For the series of spectra (left) with a moderate noise level the first 30 singular values are shown in the centered plot. The forward EFA plot is shown right. The singular values do not clearly indicate the number of absorbing species.

4.3. Spectroelectrochemical (SEC) data for a system with three chemical species

Next, we compare the two approaches to detect essential data points for UV-Vis-SEC data of some redox states of anthraquinone (AQ) [2, 26]. The measurement was carried out in a thin layer cell under argon atmosphere using a solution of 0.1 M tetrabutylammonium tetrafluoroborate in dry acetonitrile as the electrolyte. A gold mesh served as the working electrode, a platinum wire as the counter electrode, and a Ag/0.01M $AgNO_3$ electrode was used as a reference. Under these conditions, single electron reduction of AQ at $E_0 = -1.23$V leads to radical anion AQ$^-$, which can be further reduced to the dianion (AQ^{2-}) at $E_0 = -1.95$V (the equilibrium potentials E_0 have been determined in separate cyclic voltammetry experiments). During the SEC measurement, the potential was cycled once between -0.7 V and -2.7 V (sweep rate: 50mV s^{-1}). The wavelength of the absorbance peaks of all compounds are known and are used as a reference to verify our results. The signal-to-noise ratio for these UV/Vis measurements has a medium level and no matrix entry of the spectral data matrix D is close to zero. The data set consists of $k = 1000$ spectra with $n = 1397$ wavenumbers.

Fig. 15 shows the data set, the associated curve of singular values and the forward evolving factor analysis plot (EFA plot). The distribution of the largest singular values does not clearly indicate the number of chemical species. However, the knowledge of the underlying chemistry justifies to assume at least $s = 3$ species, even though the EFA plot shows more curves. We assume a moderate noise level and select $\varepsilon_C = \varepsilon_S = 0.01$ for the detection of the essential spectral information by active constraints. Fig. 16 presents the data points in the U- and V-space. We mark the data points corresponding to active constraints as determined from the dual outer polygons by red stars. Green stars indicate the vertices of the inner polygons. Additionally, the essential spectra/wavelengths corresponding to the marked data points in the mixed spectra are highlighted. We conclude that the two approaches show very similar results.

This data set indicates that although the signal-to-noise ratio is not small, a stable detection of the essential spectral

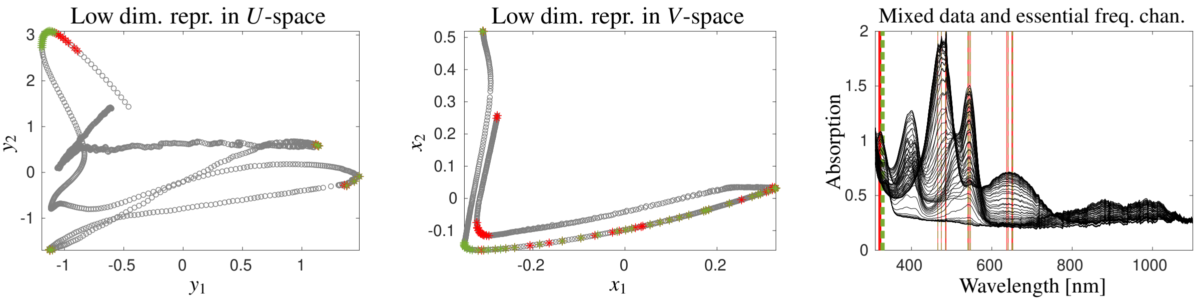

Figure 16: Spectroelectrochemical analysis of the anthraquinone system as introduced in Sec. 4.3. Left and center: All data representing points in the U- and the V-space are marked by gray circles. If such a point via duality underlies an active constraints of the outer polygons, then a red star is plotted. The convex hull of all these points forms the respective inner polygon. The vertices of the inner polygons are marked by green stars. Right: In the series of measured spectra (black) the essential wavenumbers (in general frequency channels) are marked by red vertical lines if determined by the duality-based active constraints approach and by green broken lines if they are determined by the vertices of the inner polygon. The results are comparable and indicate that both methods work well.

	dimension reduction			NMF comp. time [s]	
s	k^*	n^*	$k^*n^*/(kn)$	for D	for $\tilde{D}$
3	18	6	$3.6 \cdot 10^{-4}$	0.93	$1.00 \cdot 10^{-3}$
4	39	18	$1.7 \cdot 10^{-3}$	1.54	$1.20 \cdot 10^{-3}$
5	74	20	$6.0 \cdot 10^{-3}$	1.15	$1.40 \cdot 10^{-3}$
6	120	43	$1.6 \cdot 10^{-2}$	0.82	$3.00 \cdot 10^{-3}$

Table 1: The table lists the numbers of essential spectra k^*, essential frequency channels n^* for $s = 3, 4, 5, 6$ and the ratios of the essential data dimensions related to the full data set dimensions. The two right columns contain the computation times for the routine nnmf in MatLab applied to the original data $D \in \mathbb{R}^{k\times n}$ and the reduced data $\tilde{D} \in \mathbb{R}^{k^*\times n^*}$ for the Raman hyperspectral image data introduced in Sec. 4.2.

information is possible if all signals are significantly different from zero. Both approaches (by vertices of the inner polyhedron and by active constraints of the outer polyhedron) are suitable for this data. This finding is especially important for data with a relatively large number of chemical species, for example with $s \geq 6$, since then the computational costs to detect the active constraints strongly increases. The only significant difference can be found in the left lower corner of the V-space plot around $(x_1, x_2) \approx (-0.31, -0.11)$, namely for the spectra with indexes 845–870. This data set is also used in the final outlook section for computing a pure component decomposition based on the reduced data including only its essential parts. This reduced-data-based factorization is then used to generate a factorization of the full, original data set.

4.4. *Outlook to a dimension reduction*

Next we demonstrate the essential spectral information reduction for the spectroelectrochemical data set as introduced in Sec. 4.3; we also refer to the idea in [6]. Therefore, we define the submatrix $\tilde{D} = D(\mathrm{ess}_{C,3}, \mathrm{ess}_{S,3})$ of D which extracts from D the rows according to the lists of essential spectra indexes $\mathrm{ess}_{C,3}$ and the list of essential wavenumbers $\mathrm{ess}_{S,3}$. The numbers of indexes (or active constraints) are $k^* = 66$ spectra and $n^* = 25$. Related to the original dimensions $k = 1000$ and $n = 1397$ the information is reduced by the factor $kn/(k^*n^*) \approx 846$. Then the pure component factors underlying $\tilde{D}$ are determined. The results are shown in Fig. 17 as broken curves with respect to index subsets. The decomposition of the reduced data can be extended (SVD-based prolongated) to the complete data. A comparison with the results of a pure component decomposition for the original data shows minor deviations for the anthraquinone radical anion species.

The resulting dimension reduction has a direct impact on the computational costs for determining pure component factorizations for this data. For the spectroelectrochemical data set, see Sec. 4.3, the computation time to run the routine nnmf in MatLab is $7.28 \cdot 10^{-1}$ s for the original data $\tilde{D} \in \mathbb{R}^{1000\times1397}$ and $1.80 \cdot 10^{-3}$ s for the reduced data set $\tilde{D} \in \mathbb{R}^{66\times25}$. The same number $s = 3$ of species is used in these two cases. For the Raman hyperspectral image data introduced in Sec. 4.2 we observe similar results. Table 1 lists the reduced dimensions, the dimension reduction ratios together with the computation times needed for a nonnegative matrix factorization (NMF). All these numbers are given under the assumption of $s \in \{3, \dots, 6\}$ species. The computations were done on a single core of a standard PC with a 3.4 GHz Intel processor and 16GB RAM. The computation times are averaged over a number of 50 runs under MatLab R2018a.

5. Conclusion

A general challenge of data sciences in today's digitized world is to filter out the important parts from high-dimensional data for a subsequent dimension reduction. Transferred to MCR analyses, one is often faced with high-dimensional measured data for which the computation of a pure component decomposition is the last step of an information filtering. For this purpose, it appears advantageous to reduce the measurement data to its essential parts before applying an MCR analysis. A reliable identification of essential parts of the given data matrix appears to be necessary for a stable reduction of the data to their essential parts.

Regardless of the methodological approach, it remains to be stated that the extraction of essential or relevant spectral information from spectral mixture data is closely related to the underlying geometry of the nonnegative matrix factorization problem. The vertices of the inner polyhedron or equivalently the facets of the dual outer polyhedron are the decisive quantities which determine the feasible pure component factorizations. These fundamental relations also apply in a weakened form to noisy, experimentally gained spectral data. Then the duality-based active constraint approach appears to be a reasonable alternative to the inner-polyhedron approach. The essential information comprises only a small part (as demonstrated up to the per thousand range) of the original data dimensions. However, ignoring all the redundant, non-structure determining parts of the spectral data may not always be recommended. It is expected that redundancy can stabilize MCR factorizations, especially for a non-negligible noise level. Then a medium-sized spectral data matrix comprising the essential spectral information plus some redundant parts appears to be most reliable for MCR analyses of spectroscopic mixture data.

6. Acknowledgement

CR acknowledges funding from the project ANR-21-CE29-0007, Agence Nationale de la Recherche.

References

[1] J. J. Andrew, M. A. Browne, I. E. Clark, T. M. Hancewicz, and A. J. Millichope. Raman Imaging of Emulsion Systems. *Appl. Spectrosc.*, 52(6):790–796, 1998.

[2] A. Babaei, P.A. Connor, A. J. McQuillan, and S. Umapathy. Uv-visible spectrooelectrochemistry of the reduction products of anthraquinone in dimethylformamide solutions: an advanced undergraduate experiment. *J Chem. Educ.*, 74(10):1200, 1997.

[3] 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.

[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] M. Ghaffari, N. Omidikia, and C. Ruckebusch. Essential spectral pixels for multivariate curve resolution of chemical images. *Anal. Chem.*, 91(17):10943–10948, 2019.

[6] M. Ghaffari, N. Omidikia, and C. Ruckebusch. Joint selection of essential pixels and essential variables across hyperspectral images. *Anal. Chim. Acta*, 1141:36–46, 2021.

[7] 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.

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

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

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

[11] S. Hugelier, S. Piqueras, C. Bedia, A. de Juan, and C. Ruckebusch. Application of a sparseness constraint in multivariate curve resolution ? Alternating least squares. *Anal. Chim. Acta*, 1000:100–108, 2018.

[12] C. Kubis, M. Sawall, A. Block, K. Neymeyr, R. Ludwig, A. Börner, and D. Selent. An operando FTIR spectroscopic and kinetic study of carbon monoxide pressure influence on rhodium-catalyzed olefin hydroformylation. *Chem.-Eur. J.*, 20(37):11921–11931, 2014.

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

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

[15] A. Meister. Estimation of component spectra by the principal components method. *Anal. Chim. Acta*, 161:149–161, 1984.

[16] 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.

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

[18] C. Ruckebusch, R. Vitale, M. Ghaffari, S. Hugelier, and N. Omidikia. Perspective on essential information in multivariate curve resolution. *Trend Anal. Chem.*, 132:116044, 2020.

[19] 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.

[20] 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.

[21] M. Sawall, C. Kubis, H. Schröder, D. Meinhardt, D. Selent, R. Franke, A. Brächer, A. Börner, and K. Neymeyr. Multivariate curve resolutions methods and the design of experiments. *J. Chemom.*, 32(6):e3012, 2019.

[22] M. Sawall, A. Moog, C. Kubis, H. Schröder, D. Selent, R. Franke, A. Brächer, A. Börner, and K. Neymeyr. Simultaneous construction of dual Borgen plots. II: Algorithmic enhancement for applications to noisy spectral data. *J. Chemom.*, 32:e3012, 2018.

[23] 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.

[24] 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.

[25] 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.

[26] M. Shamsipur, B. Hemmateenejad, A. Babaei, and L. Faraj-Sharabiani. Use of multivariate curve resolution analysis in the spectroelectrochemistry of 9, 10-anthraquinone reduction in dimethylformamide solution. *J Electroanal. Chem.*, 570(2):227–234, 2004.

[27] 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.

[28] M. Vosough, C. Mason, R. Tauler, M. Jalali-Heravi, and M. Maeder. On rotational ambiguity in model-free analyses of multivariate data. *J. Chemom.*, 20(6-7):302–310, 2006.

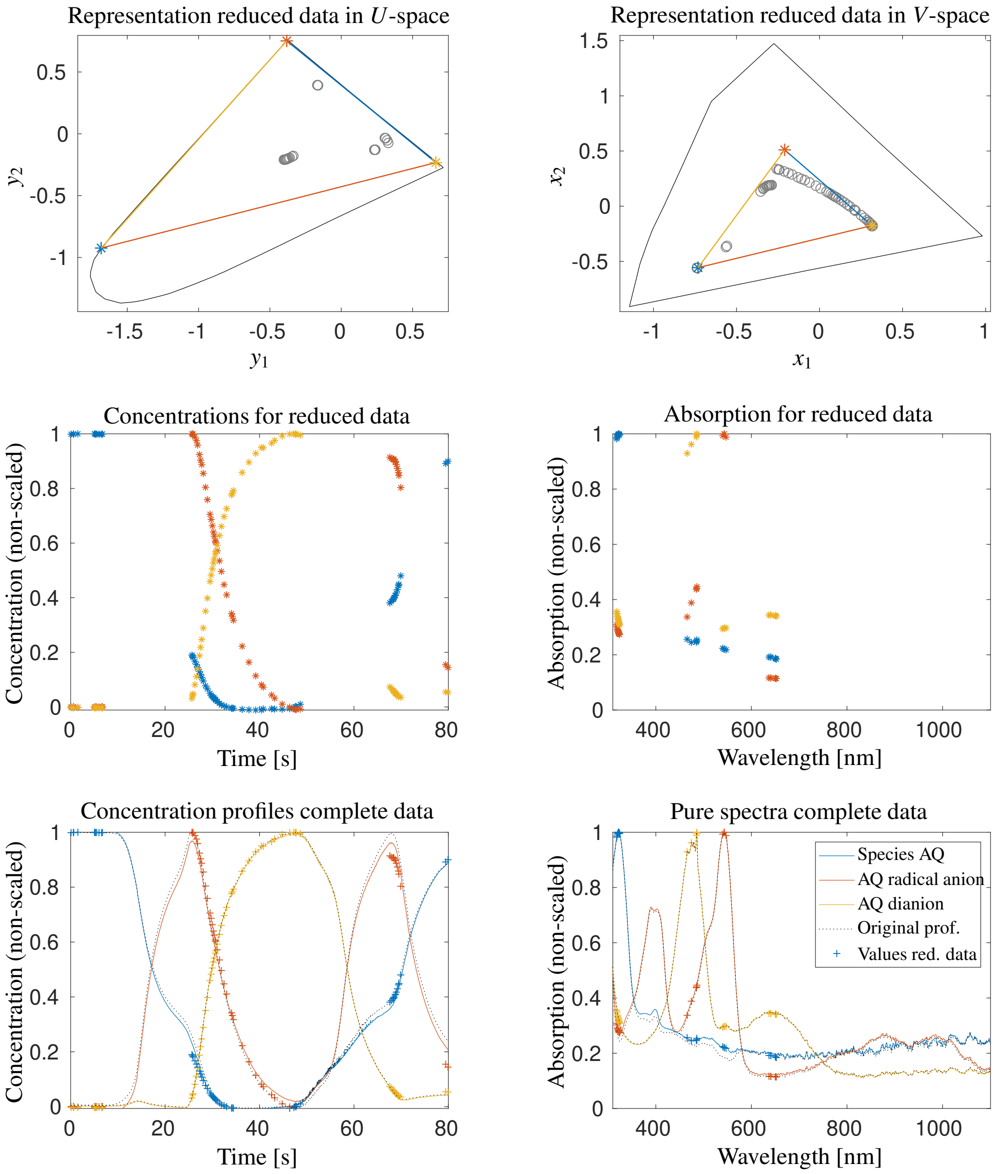

Figure 17: Reduction of the anthraquinone (AQ) spectroelectrochemical data, see Sec. 4.3, to its essential parts. First row: Low-dimensional data representing vectors in the U- and V-space for the reduced data set $\tilde{D} \in \mathbb{R}^{66\times 25}$. Second row: The associated concentration and absorption values with respect to the grids of essential indexes. Third row: The discrete profiles (colored stars) from above are underlaid with the concentration profiles and pure component spectra (colored lines) as computed for the complete data. The results of a direct factorization of the complete data D are represented by black dotted lines and the values from the 2nd row are marked by black pluses (the results based on $\tilde{D}$.