A frequency channel scanning method for the Peak Group Analysis

Lukas Prestin[a,b], Mathias Sawall[a], Esha Suresh Babu[b], Chunhong Wei[b], Chistoph Kubis[b], Klaus Neymeyr[a,b]

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

Abstract

Peak Group Analysis (PGA) is a Multivariate Curve Resolution (MCR) algorithm that extracts pure component spectra and concentration profiles from IR and Raman data with high spectral selectivity. PGA solves constrained convex minimization problems (coupled with non-convex soft constraints) exclusively for selected frequency channels that correspond to peak positions in the mixture spectra.

In this work, we introduce a numerically efficient frequency channel scanning algorithm for PGA. The proposed approach solves the minimization problems across all frequency channels, avoiding the often-problematic peak detection step. The resulting residuals reliably identify the relevant pure component spectra. Additionally, the calculated residuals across non-peak channels provides useful insights into the feasibility and factor ambiguity of the solution.

Keywords: multivariate curve resolution, peak group analysis, nonnegative matrix factorization, pure component decomposition

1. Introduction

Multivariate curve resolution (MCR) techniques serve to analyze spectral mixture data and extract information about the underlying pure chemical components. In spectroscopic applications such as Fourier-transform infrared (FT-IR) and Raman spectroscopy, measured spectra typically contain locally overlapping contributions from multiple chemical species. The goal of MCR methods is to recover both the pure component spectra and their corresponding concentration profiles from the mixed data.

For a spectral dataset represented by the matrix $D \in \mathbb{R}^{k\times n}$, containing k mixture spectra measured across n frequency channels, the decomposition is typically modeled according to the Lambert-Beer law

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

where $S \in \mathbb{R}^{n\times z}$ contains the pure component spectra, $C \in \mathbb{R}^{k\times z}$ represents the corresponding concentration profiles, and E denotes the residual error matrix. The natural number z denotes the number of chemical species.

Peak Group Analysis (PGA) is a specialized MCR approach for determining pure component spectra in FT-IR and Raman spectral data. It exploits spectral selectivity in specific frequency intervals and evaluates candidate spectra using an objective function built to isolate chemically meaningful pure component spectra. PGA extracts these candidate spectra by solving optimization problems. The core idea is to minimize the Euclidean norm of a potential spectrum Vt (being a linear combination of right singular vectors of D stored column-wise in V) subject to constraints: the signal at a selected target channel ℓ is forced to equal 1 ($(Vt)_\ell = 1$), and the spectrum is constrained to be nonnegative ($Vt \geq 0$). Additional (soft) constraints can be added. The optimization isolates the target component while suppressing contributions from interfering species [15, 16].

Despite its central role, the exact effect of this channel-wise optimization strategy on the performance of PGA has not been systematically studied. In this work, we investigate the impact of this approach by introducing a frequency channel scanning algorithm that applies the channel-targeted constraint $(Vt)_\ell = 1$ across the entire frequency range.

1.1. Overview

Sec. 2 introduces the basic idea of Peak Group Analysis with a focus on the objective function used to determine chemically meaningful pure component spectra and the required constraints on the candidate spectra. Sec. 3 presents the new systematic frequency chan-

nel scanning algorithm and its results for Datasets 1 and 2. In Sec. 4 these results are analyzed using low-dimensional representations to provide an improved understanding of the PGA algorithm.

1.2. Datasets

The method is illustrated using two experimental datasets from catalysis research at the Leibniz Institute for Catalysis, Rostock, Germany, to demonstrate the behavior of the algorithm and its potential usefulness as a spectral analysis tool.

Dataset 1. *The FT-IR spectroscopic dataset represents the formation of inactive polynuclear rhodium carbonyl complexes, using $Rh(acac)(CO)_2$ as a precatalyst for alkene hydroformylation in the absence of phosphorus ligand [7]. The time-resolved in situ spectroscopic experiment was conducted at* $100°C$ *under 20 bar of synthesis gas ($CO/H_2 = 1:1$) with an initial concentration of the starting material of* $1 \cdot 10^{-3}$ *mol/L in n-dodecane as solvent. Throughout the reaction, $Rh(acac)(CO)_2$ is progressively converted into $Rh_4(CO)_{12}$ and subsequently $Rh_6(CO)_{16}$ in a consecutive manner.*

The resulting dataset consists of 208 *spectra with* 1453 *data channels (wavenumber values) within the range of* $1800\,\mathrm{cm}^{-1}$ *to* $2150\,\mathrm{cm}^{-1}$. *The absorption spectra are characterized by sharp, partially overlapping peaks with relatively high signal-to-noise ratios stemming from the infrared active vibrational modes of the three named components. Acetylacetone (acacH) is also released during this process, but its infrared bands do not absorb the respective spectral range. Because the dataset includes small negative entries introduced by background subtraction, it serves for studying the impacts of noise and data processing artefacts. The spectra of this three-component system can be represented in the two-dimensional V-space, see Sec. 1.3.*

Dataset 2. *The FT-IR dataset was collected during a time resolved in-situ spectroscopic squential dosage experiment in the framework of a component identification analysis at* $120°C$ *und* 20 *bar of synthesis gas with n-dodecane as solvent. The process covers the sequential addition of first diphenyl carbonate DPC as internal standard, followed by $Rh(acac)(CO)_2$ and then synthesis gas CO/H_2 which leads to the formation of $Rh_4(CO)_{12}$ and $Rh_6(CO)_{16}$. The subsequent addition of the P-ligand tris-(2,4di-tert.-butylphenyl)phosphite(TDTBPP) then leads to the formation of $HRh(CO)_3L$ (L=TDTBPP), which is an active hydroformylation catalyst. Basic spectroscopic characterization on this catalyst can be found in the literature [5].*

The obtained data series comprised of 1275 *spectra on* 3319 *data channels (wavenumber values) within the spectral range of* $1500\,\mathrm{cm}^{-1}$ *to* $2300\,\mathrm{cm}^{-1}$. *This large dataset, with its time-dependant spectral contributions (multiple peaks with overlapping absorptions) from seven components, is well suited for evaluating the method presented in this paper.*

1.3. Low-dimensional representations in the V- and U-space

The common way is to represent the results and intermediate steps of PGA by means of low-dimensional representations in V-space and U-space defined by the right and left singular vectors from a singular value decomposition (SVD) of D. The area of feasible solutions (AFS, [1, 3, 18]) is also represented in this space and provides the basis for comparing the results.

The measured spectra as well as vectors calculated as potential pure component spectra can be represented in a low-dimensional form by using their coefficients with respect to the basis of right singular vectors. In the V-space, a candidate spectrum

$$s = Vt$$

is represented by a z-dimensional vector t if z singular vectors are considered. It is well known (Perron-Frobenius theory) that the first component of t is nonzero. This justifies applying a scaling with $1/t_1$ such that

$$\frac{1}{t_1} t = \left(1, \frac{t_2}{t_1}, \ldots, \frac{t_z}{t_1}\right)^T .$$

Omitting the leading 1, a $(z-1)$-dimensional representation is possible. For Dataset 1 with $z = 3$ this enables the representation in two-dimensional space using x_1 and x_2 to describe $V(1, x_1, x_2)^T$; see Figs. 2, 3, 5 and 9.

In a similar manner, concentration profiles and temporal evolutions of specific frequency channels can be plotted in the U-space. A concentration profile

$$c = U\Sigma r$$

has a $(z-1)$-dimensional representation with the scaled expansion coefficients $(1, \frac{r_2}{r_1}, \ldots, \frac{r_z}{r_1})$ and where the leading 1 can be skipped. A U-space plot is shown in Fig. 7.

1.4. Notation

The following notation is used in the paper.

D	$k \times n$ spectral data matrix
C	$k \times z$ concentration matrix
S	$n \times z$ spectra matrix
$U\Sigma V^T$	truncated singular value decomposition of D
z	number of largest singular values used in truncated SVD
s	candidate pure component spectra
t	linear coefficients such that $s = Vt$
ℓ	target frequency channel index
$\mathcal{L}$	set of detected peak center indices
F	objective function

2. Peak Group Analysis

PGA is an MCR method designed to identify the pure component spectra from time-resolved spectral mixtures. The procedure targets a specific peak in the mixture data and reconstructs the full pure component spectrum anchored to that peak position. Reconstructing a signal based on a localized peak effectively extracts the entire correlated spectral signature of that specific chemical species.

This works particularly well for selective peaks or even just selective frequency channels [13]. A region is selective if only a single chemical component exhibits significant absorption within it. Since selective regions frequently occur in FT-IR and Raman spectroscopy, PGA and similar routines, such as band target entropy minimization (BTEM [23, 25]), can be used successfully with such data, even in the presence of noise.

2.1. PGA version history

PGA was initially introduced in [15], where frequency channel intervals I containing peaks were manually selected to determine the underlying pure component spectra. An automated peak detection framework was later implemented [16], offering an alternative to manual selection. Automated Peak Group Analysis (autoPGA) detects peak centers using a combination of criteria, such as second order derivatives or changes along the time axis. AutoPGA significantly accelerates the chemometric analysis and removes dependency on manual input.

However, both peak detection techniques are susceptible to overlooking essential peaks, particularly those belonging to intermediate chemical species present at low concentrations. The frequency channel scanning approach presented in this paper does not require an *a priori* peak detection and thus reduces the manual input.

2.2. How PGA works

Like many established MCR routines [4, 8, 9, 10], PGA uses the singular value decomposition of the spectral data matrix D

$$D = U\Sigma V^T.$$

The columns of the orthogonal matrices U and V serve as bases for the concentration profiles and pure component spectra, respectively. The diagonal matrix Σ contains the singular values, providing further information on the required number of used singular vectors [21].

PGA uses soft constraints to determine the pure component spectra by numerical optimization. The autoPGA algorithm within the FACPACK software package [20] comprises three main steps:

1. First the selection of a set $\mathcal{L}$ of peak center indices ℓ is identified, see Fig. 1. This is done either via automated pattern recognition [16] or through the manual selection of intervals I.
2. For each selected peak center $\ell \in \mathcal{L}$, a candidate pure component spectrum is constructed as a linear combination of the first z right singular vectors. The coefficients are determined by minimizing a specific objective function under the strict channel (normalization) constraint $Vz\,|_\ell = 1$.
3. Finally, a covariance analysis filters the resulting spectra to skip repeated detections.

Step 2 is the computational core of the algorithm. Potential pure component spectra are extracted using a soft modeling approach controlled by an objective function. This function penalizes negative absorptions, large Euclidean norms, and other undesirable spectral characteristics. The behavior of the objective function can be fine-tuned for specific chemical applications via weighting factors.

2.3. The objective function

Spectra are constructed as linear combinations of the first z right singular vectors $V_{:,j}$

$$s = \sum_{j=1}^{z} V_{:,j} t_j = Vt \tag{2}$$

where the coefficient vector t is to be determined. The truncation parameter z is chosen based on the expected

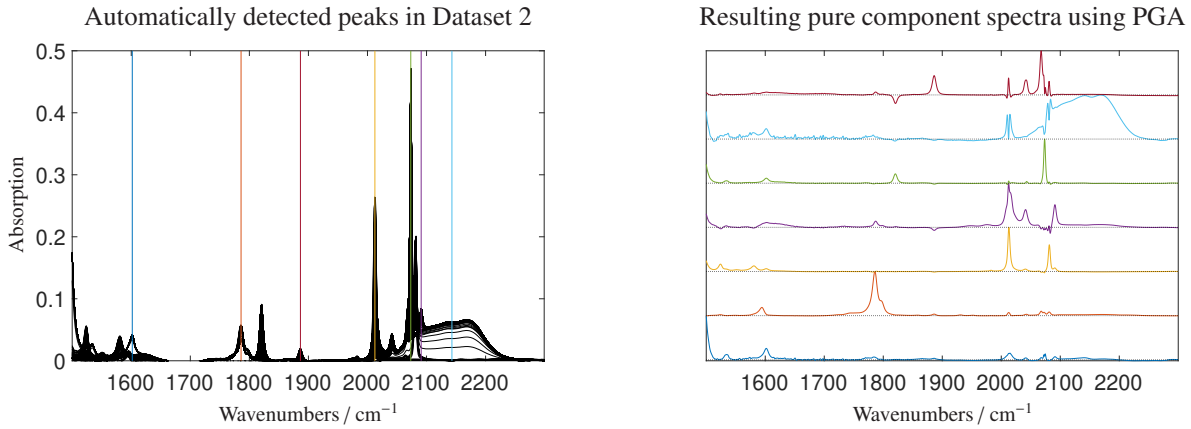

Figure 1: Results of PGA for Dataset 2 obtained from a sequential dosage experiment. Left: Spectral mixture dataset with automatically detected peaks highlighted. Right: Resulting pure component spectra obtained from the PGA calculations. Each selected peak is used in the optimization of an objective function to generate one possible pure component spectrum, shown in the corresponding color.

number of pure chemical species and the magnitude of the singular values. To obtain chemically meaningful solutions, candidate spectra are evaluated against several mathematical criteria. Deviations from certain expectations result in high values of the objective function. This allows us to identify profiles that are nonnegative, smooth and have a small Euclidean norm. The objective function $F : \mathbb{R}^n \to \mathbb{R}$ reads

$$F(s) = \|s\|_2^2 + \omega_0 \|\min(s, 0)\|_2^2 + \sum_{j=1}^{4} \omega_j f_j(s) \qquad (3)$$

with weighting factors $\omega_j \geq 0$ and penalty functions $f_j : \mathbb{R}^n \to \mathbb{R}$, $j = 1, \ldots, 4$. F is described in a similar way in [15, 16] aside from the numbering of the constraint functions. Most important in this objective function is the Euclidean norm of considered spectra to favor spectra with sharp, isolated peaks. The nonnegativity constraint is scaled by $\omega_0 > 0$. This constraint can be adjusted to tolerate small negative entries greater than $-\varepsilon$ (with a control parameter $\varepsilon > 0$) to account for data with an imperfect baseline correction and noise.

The additional constraints in (3) can be used to fine-tune the reconstructed spectra acording to specific assumptions. Smoothness in the candidate spectra can be enforced using a discrete second derivative approximation

$$f_1(s) = \sum_{i=2}^{n-1} \left(\frac{s_{i-1} - 2s_i + s_{i+1}}{\Delta^2} \right)^2$$

where Δ denotes the spacing between consecutive wave numbers. If a peak is manually selected over an interval I ,see [15], the local reconstruction error

$$f_2(s) = \left\| \Sigma \left(V_{:,j}^T - V_{:,j}^T \frac{s_I^T s_I}{\|s_I\|_2^2} \right) \right\|_F^2$$

can be added to the objective function. After determining the first few pure component spectra $s^{(1)}, \ldots, s^{(m)}$ in PGA,

$$f_3(s) = \sum_{j=1}^{m} (s^T s^{(j)})^2$$

can be applied to favor small covariances, thus achieving low similarity to these already computed spectra. Prior knowledge of the dataset can be introduced to the optimization by using an expected pure component spectrum $\tilde{s}$ measured in isolation in

$$f_4(s) = \sum_{i=1}^{n} \left(\alpha \tilde{s}_i - s_i \right)^2, \quad \text{with } \alpha = \frac{\tilde{s}^T s}{\|\tilde{s}\|_2^2}$$

to achieve high similarity to this reference spectrum. The value of α is chosen to optimally scale the spectrum for such a comparison.

All constraint functions used in F are sums of squares, it is therefore efficient in the implementation to approach the optimization as a nonlinear least squares problem, that can be solved using established methods [2, 11].

The global minimum of this objective function occurs at the chemically meaningless trivial $s = 0$, resulting in

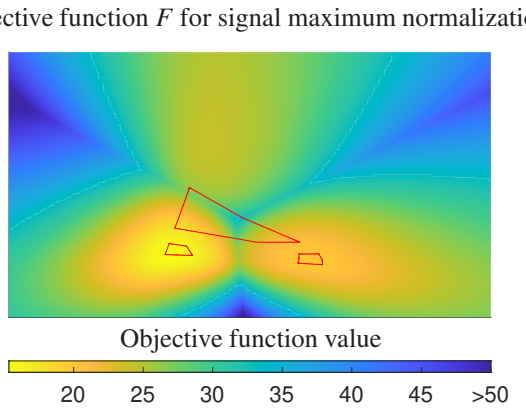

Figure 2: V-space plot depicting the objective function for candidate spectra in low-dimensional representation for Dataset 1. The red outlines denote the area of feasible solutions (AFS), containing all valid combinations of pure component spectra for nonnegative matrix factorizations. The objective function surface exhibits several local minima (yellow), resulting from the $\max(s) = 1$ scaling, which complicates the identification of pure components by introducing multiple convergence points for optimization algorithms.

$F(0) = 0$. To avoid this result, a normalization constraint for the considered candidate spectra is required. Here a commonly used normalization approach is tested and found to be inadequate in Sec. 2.4. Subsequently, the peak-specific normalization used in PGA is introduced in Sec. 2.5.

2.4. The problem with signal maximum normalization

A commonly used normalization method is to scale the maximum of a considered spectrum $s = Vt$ to 1, see [6, 8]. However, this normalization leads to non-convexity of the objective function F and thus complicates the optimization. Fig. 2 illustrates this for Dataset 1 in terms of the area of feasible solutions (AFS,[1, 3]). For every candidate spectrum $s = Vt$ with $t = (1, x_1, x_2)^T$ this V-space plot shows the objective function using the internal signal maximum normalization $F(s/\max(s))$. The resulting objective function surface is complex and contains multiple local minima associated with different chemical species. Depending on the weighting factors ω_j, finding the true pure component spectra requires extensive global optimization or specific starting points.

2.5. Peak-specific normalization

To avoid the challenges produced by this global optimization, PGA uses a peak-specific normalization approach. For each peak detected in the first step, the candidate spectrum $s = Vt$ is scaled such that its value at the specific frequency channel index of the peak center ℓ equals 1, thus the optimization problem reads

$$\min_{t\in\mathbb{R}^z} F\left(\frac{Vt}{(Vt)_\ell}\right) \tag{4}$$

where $\ell \in \mathcal{L}$ represents a channel index from the set of identified peak centers or $\ell = \operatorname{argmax}(Vt)_I$ for a manually selected interval I containing a peak. This normalization step is crucial because it directly influences the norm constraint within the objective function and therefore affects which candidate spectra are selected, favoring those with high absorption in the selected frequency channel [13, 15].

Fig. 3 illustrates the impact of this channel-specific normalization on Dataset 1 in terms of the AFS. In contrast to the non-convex surface with multiple local minima shown in Fig. 2, the peak-specific normalization effectively isolates a single minimum for each chemical component. This approach ensures reliable convergence of the optimization to the pure component spectrum associated with the selected frequency channel.

The procedure of PGA with frequency channel normalization is outlined in Algorithm 1. In the more recent autoPGA [16], the manual selection of intervals containing peaks is replaced by automated peak detection. In this preliminary step, routines return several single frequency channels as the assumed centers of the detected peaks. Then, PGA is applied to these frequencies simultaneously. A correlation analysis is applied afterwards to determine repeated detections of the same pure component spectra.

Algorithm 1 Automated Peak Group Analysis

Input: D, weights $\omega_0, \ldots, \omega_4$, no. singular vectors z

Step 1 - peak selection:

select set of peak centers in frequency indices $\ell \in \mathcal{L}$

calculate SVD $D = U\Sigma V^T$

Step 2 - objective function optimization:

for each $\ell \in \mathcal{L}$ **do**

solve the optimization problem to compute $s^{(\ell)}$, thus

$$s^{(\ell)} = \operatorname*{argmin}_{s=Vt,\, s_\ell=1} F(s)$$

end for

Step 3 - final selection: determine similar resulting spectra $s^{(\ell)}$ using covariance analysis

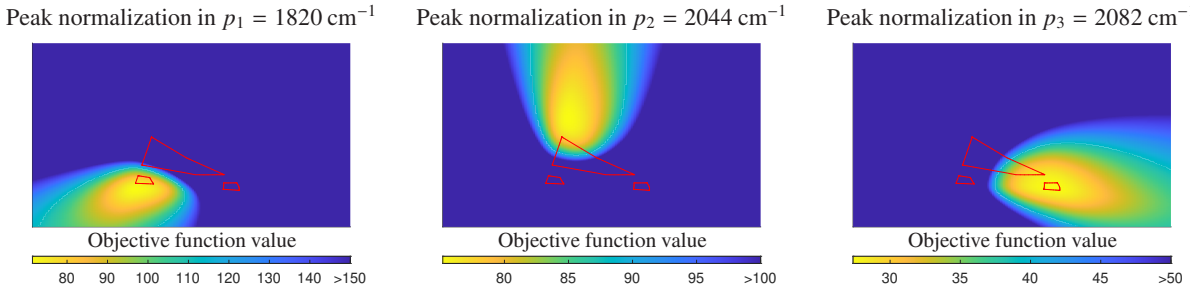

Figure 3: V-space plots depicting objective function values depending on the linear coefficients of candidate spectra for Dataset 1. The dark blue areas represent high objective function values and are assigned to spectra that are not chemically meaningful. The best spectra according to the selected peak and objective function are shown in yellow. The AFS regions are marked in red for comparison. The normalization approach based on the selection of peaks is applied within PGA. The resulting objective function values for low-dimensional representations of considered spectra show a clear distinction with smallest values for spectra containing absorption in the chosen peak. The selection of peaks in frequency channels p_1, p_2 and p_3 leads to three different priorities for the objective function, thus allowing for three pure component spectra to be detected.

2.6. Correlation analysis

Because multiple peak centers may originate from the same chemical species, duplicate pure component spectra are often determined in the optimization. To address this, a covariance analysis is applied to filter the resulting set of spectra $s^{(\ell)}$. Repeated detections are determined by grouping together results $s^{(\ell)}, s^{(\tilde{\ell})}$, where the cosine of the acute angle [12, 22]

$$\cos\theta = \frac{\sum_i s_i^{(\ell)} s_i^{(\tilde{\ell})}}{\left\|s^{(\ell)}\right\| \left\|s^{(\tilde{\ell})}\right\|} \tag{5}$$

exceeds a threshold of 0.9, indicating a high similarity between these spectra. This way, repeated detections of the same components are prevented.

3. Systematic Frequency channel scanning

As an alternative approach within PGA, we replace the peak selection respectively the automatic peak detection and implement a systematic frequency channel scanning method. Rather than restricting the normalization constraint to pre-selected peak centers, this approach iterates through every available frequency channel in the dataset. This scanning exploration removes possible selection biases of the a priori selection of peaks and reveals the complete optimization results under the applied constraint. Thus, minor and intermediate chemical species that exist at low concentrations and might be overlooked during the peak detection in autoPGA can be determined.

3.1. The algorithm

The proposed frequency channel scanning algorithm replaces the selection of peaks in the classical PGA respectively the first step of autoPGA by systematically iterating through all available frequency channels. For every channel $\ell = 1, \ldots, n$ the corresponding optimization problem (4) is solved. If a pure component exhibits a peak at the ℓth frequency channel, the objective function is expected to achieve a prominent local minimum and the optimization will converge to the corresponding pure component spectrum. The resulting spectra are subsequently compared based on their associated objective function values. Here, the objective function described in (3) is considered using the same weighting factors for all frequency channels, allowing for an unbiased comparison of results.

The normalization according to (4) ensures that the currently considered frequency channel is explicitly accounted for during the detection of possible pure component spectra. This algorithm identifies the linear combination of first right singular vectors that minimizes the objective function, thus providing the spectrum that is most chemically meaningful according to the selected weightings of constraints. The corresponding minimal objective function values can be plotted across the spectral range to visualize the most promising channels to use in PGA, as illustrated in red in Fig. 4 and 5.

Although frequency channel scanning requires solving one nonlinear optimization problem per channel, the computational effort is managed efficiently using a warm start approach. Using the optimization results from the previous frequency channel as starting values for the subsequent step often only minor adjustments are required, reducing the runtime significantly. This is, of course, based on the assumption that the resulting spectra change continuously with small changes in frequency channels. Rather than solving n independent non-linear optimization problems, the runtime for successive frequency channels with similar optimization

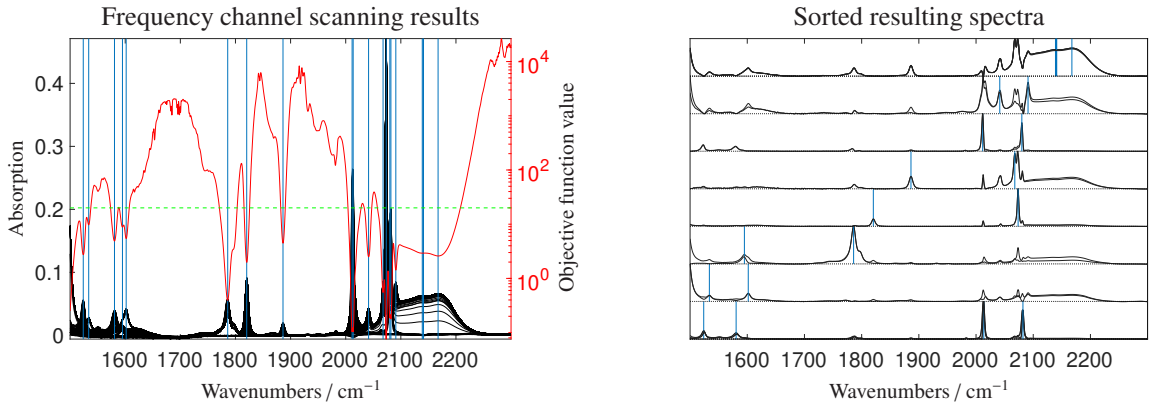

Figure 4: Frequency channel scanning results for Dataset 2. Left: Frequency channel scanning minimizes the objective function for each frequency channel considered for normalization, here marked in red. The local minima in these values below a threshold $\delta = 20$ correspond to peaks in the measured data, that can be used to acquire pure component spectra. These are marked as blue vertical lines in the corresponding frequency channels. Right: Frequency channel scanning results in the detection of 20 spectra, that are sorted into eight sets of similar spectra using the acute angle described in (5).

Algorithm 2 Frequency channel scanning

Input: D, weights $\omega_0, \ldots, \omega_4$, no. singular vectors z, threshold δ

compute SVD $D = U\Sigma V^T$

for $\ell = 1$ **to** n **do**

solve the optimization problem to compute $s^{(\ell)}$, thus

$$s^{(\ell)} = \underset{s=Vt,\, s_\ell=1}{\operatorname{argmin}} F(s)$$

(if $\ell > 1$, use $s^{(\ell-1)}$ as starting point)

end for

detect spectra $s^{(\ell)}$ with locally minimal objective function values $F(s^{(\ell)}) < \delta$

determine similar resulting spectra $s^{(\ell)}$ using covariance analysis

problems is reduced drastically.

3.2. Evaluation of results

Frequency channel scanning provides the smallest achievable objective function value F and the corresponding spectra $s^{(\ell)}, \ell = 1, \ldots, n$ under the constraints $s^{(\ell)}_\ell = 1$. These results can be used to gain insights into the spectroscopically monitored processes and improve our understanding of the functionality of PGA.

To isolate valid candidate spectra from worse results associated with high residuals, a threshold δ is applied. Local minima of the objective function, that fall below this threshold, i.e., $F(s^{(\ell)}) < \delta$ correspond to valid pure component spectra.

The minimal objective function values are illustrated in red in Fig. 4 in front of Dataset 2. The local minima are marked in blue and the dotted green line represents the threshold, which was chosen manually as $\delta = 20$.

The frequency channels identified in this way correspond well with the peak centers, some of which were also identified by previously used peak detection methods. However, determining these channels directly from the optimization results makes it possible to identify all wavenumbers that lead to pure component spectra. For comparison, Fig. 6 plots the frequency channel scanning results alongside the purest channels identified by the SIMPLISMA method [24] and essential frequency channels [19] from the FACPACK software package. The scanning algorithm successfully captures the purest points identified using these methods.

The prior selection of peaks in the first step of PGA is replaced by a frequency channel scanning approach that ensures that the lowest possible objective function values for the selected weighting factors are achieved. This approach is especially useful for noisy data, where choosing the correct frequency channels corresponding to peak centers might be challenging.

The objective function F in (3) is chosen to provide low values for chemically meaningful results, assuming those to have a relatively small Euclidean norm and no substantial negative entries. The resulting residuals, shown in red in Fig. 4, contain local minima associated with valid pure components, marked in blue. Frequency channels without local minima or with objective func-

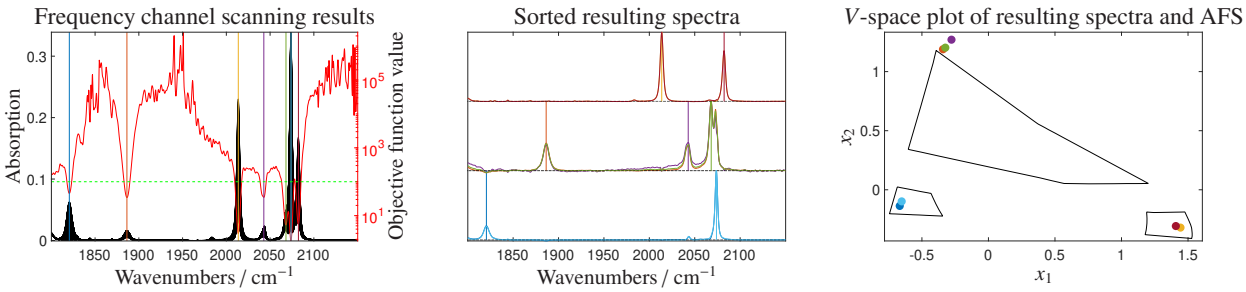

Figure 5: Frequency channel scanning results for Dataset 1. Left: The red curve represents the minimal objective function value for each frequency channel. Local minima below the threshold $\delta = 100$ are marked as vertical lines and correlate to the best resulting spectra. Center: The resulting spectra are sorted based on a covariance analysis, leading to three sets of distinct pure component spectra. Right: V-space plot of the resulting spectra showing a clear distinction of the three sets. These can each be assigned to one of the three subsets of the AFS-subsets marked in black.

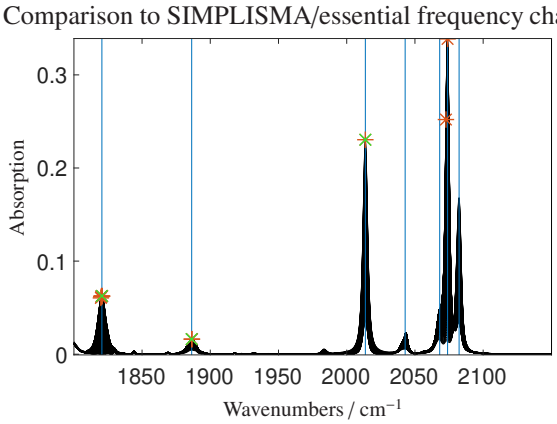

Figure 6: Frequency channel scanning results (blue) in comparison to SIMPLISMA (green) and essential frequency channels (orange). Both methods can be used to detect the purest frequency channels from Dataset 2. The wavenumbers 1820cm^{-1}, 1886cm^{-1} and 2014cm^{-1} were selected by all three methods, the peak in 2074cm^{-1} was detected by frequency channel scanning and by the essential spectral information algorithm.

tion values larger than the threshold $\delta = 20$ are not used for the determination of the pure component spectra.

3.3. Similarity analysis

Applying the similarity criterion in (5) to the channel scanning results allows us to group them to distinct chemical species. As shown in Fig. 4, the 20 resulting spectra are grouped into eight pure component spectra. For Dataset 1, the seven identified channels and the corresponding spectra are successfully assigned to the three underlying pure component spectra in Fig. 5. High similarity between two spectra is associated with small distances between the low-dimensional representations in the V-space plot.

In Fig. 7, the acute angle is used to mark all frequency channels with a high similarity to one of the three underlying pure component spectra in the corresponding color. The right plot allows for a clear distinction between these frequency channels resulting in different spectra in terms of the U-space representation.

3.4. Choice of window size

AutoPGA allows peaks to be selected in frequency intervals of different sizes, either per manual selection or in the form of individual frequency channels provided by automated peak detection methods [16]. Similarly, the frequency channel scanning approach can be extended to operate on multi-channel windows. In this case, we normalize the candidate spectra in the index interval $I = [\ell - m, \ell + m]$ in the maximum to 1. Thus, we solve the optimization problem

$$s^{(\ell)} = \operatorname*{argmin}_{s=Vt,\, \max(s(I))=1} F(s).$$

Intervals of length 1 ($m = 0$) correspond to single frequency channels and are applied in Fig. 4 and 5. Fig. 8

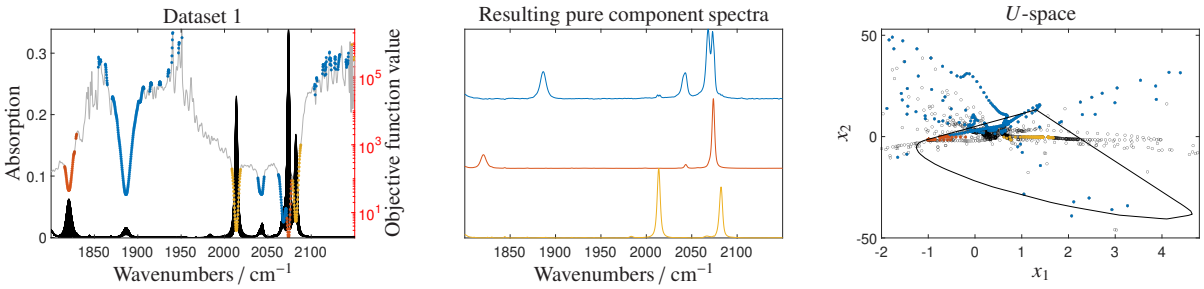

Figure 7: Frequency channel scanning results can be matched to the pure component spectra. Left: Dataset 1 with smallest possible objective function values resulting from frequency channel scanning marked in grey. The cosine of the acute angle (see (5)) between resulting spectra and a pure component exceeding 0.9 is used as a criterion to match these together. All frequency channels used in the normalization step to achieve these corresponding spectra are marked in the respective color. Middle: Pure component spectra achieved as frequency channel scanning results. Right: Low-dimensional representation of measured frequency channels in U-space plot. A clear distinction of the three sets of corresponding colors is visible.

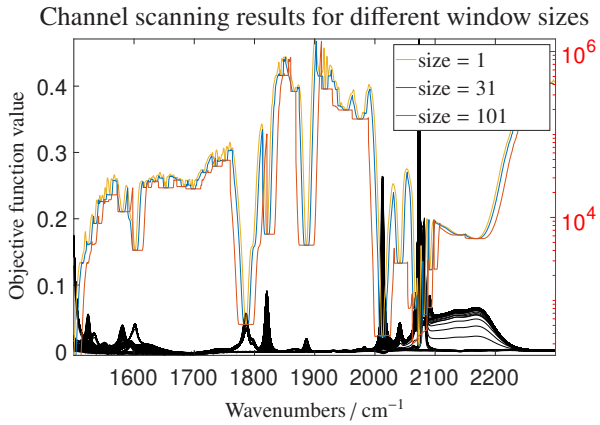

Figure 8: Frequency channel scanning results for a moving window approach with the window sizes 1 (see Fig. 4), 31 and 101. While overall shape and local minima are similar, for larger window sizes plateaus appear around local minima, as similar objective function values are achieved for all windows containing shared peaks.

shows the optimal objective function values for spectra fitted to a moving window containing 31 or 101 frequency channels, respectively.

Due to the normalization underlying the optimization, many frequency windows containing the same peak achieve a similar objective function value. Therefore, the results of this moving window scanning for $m > 0$ cannot be used as a peak detection method. The observed behavior of lowest determined objective function values for window sizes larger than 1 is similar to results for single frequency channels and shows the most promising results for peaks with higher amplitude.

4. Uncover partly overlapping peaks

Selective regions are frequency channel intervals that contain information from only one of the underlying chemical components in a mixture. PGA works particularly well in such areas. Furthermore, these regions are sufficient to compute a complete factorization by PGA [13]. In combination with the area of feasible solutions (AFS) respectively the V-space plots, the frequency channel scanning results can be used to visualize the effect of (more-or-less) selectivity on PGA.

In the following example, we focus on a small frequency interval of Dataset 1. In this window, the data contains information from three components in the form of three peaks that are close to each other and partly overlapping. We analyze the frequency channel scanning in this interval in terms of the AFS. The PGA results are plotted within the V-space and show the shift between the points representing the three pure component spectra. The resulting pure component spectra, recovered by minimizing the objective function for these three peaks, are distinct. The low-dimensional representation of these resulting spectra are shown as points with corresponding colors in the bottom right plot of Fig. 9 and can be clearly assigned to different areas of the AFS.

In contrast, the intermediate frequency channels lying between these peak centers are non-selective, where the measured signal is a linear combination of multiple components. The spectra that are optimal according to the minimization of the objective function for the frequency channel in this range resemble a transition between the resulting spectra for ℓ_1 and ℓ_2 or ℓ_2 and ℓ_3, respectively. In the V-space, the low-dimensional representations of these intermediate results trace a continuous path between the results for the peak centers. This

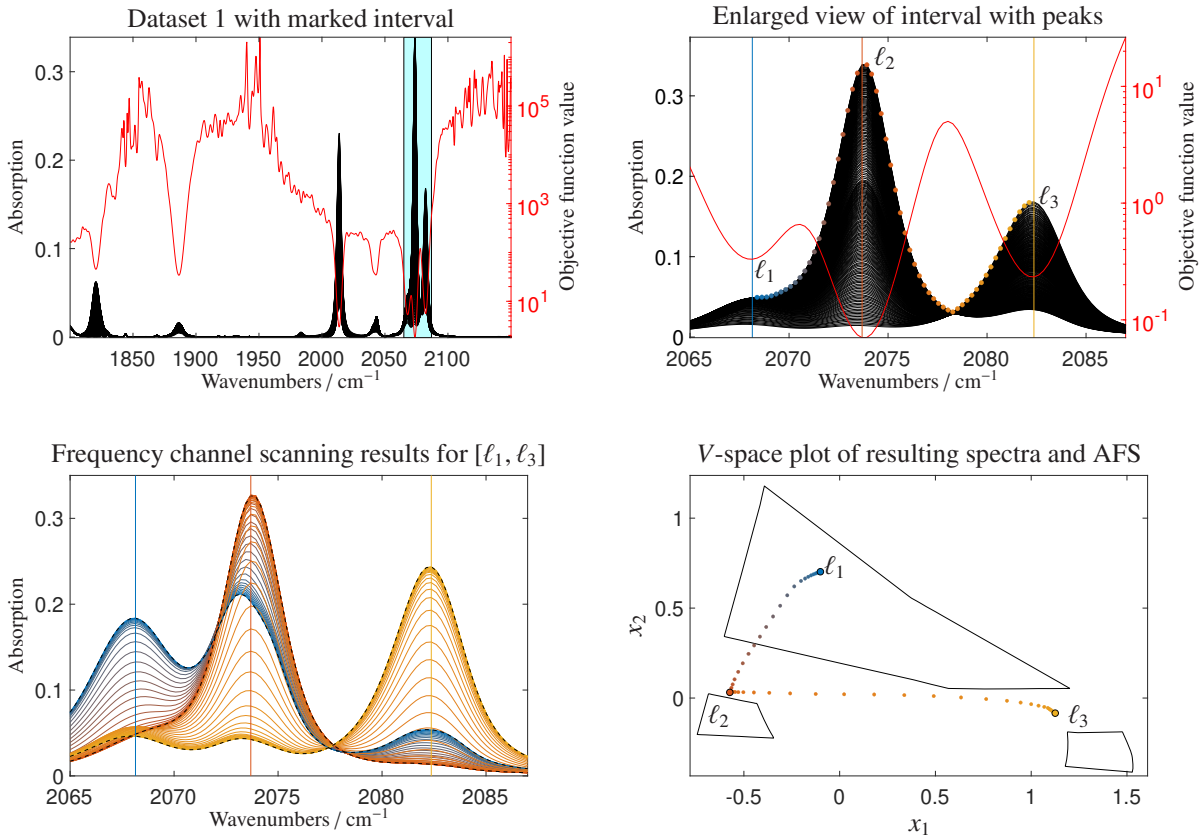

Figure 9: A frequency interval of Dataset 1 containing three peaks at the indexes ℓ_1, ℓ_2, ℓ_3 belonging to different pure component spectra. Upper left: The measured spectra with the associated optimal objective function values for the frequency channels on the interval. Upper right: Enlarged plot of interval with marked peak centers. Lower left: Resulting spectra providing the lowest determined objective function values for indexes from ℓ_1 to ℓ_3. Clear transitions from the resulting spectrum achieved using normalization in ℓ_1 (blue) to ℓ_2 (orange) can be observed. The same can be seen from the second center to ℓ_3 marked in ochre. Lower right: The AFS in the low-dimensional representation including the representers of the PGA results for the index interval $[\ell_1, \ell_3]$. Here the transition of results for frequency channels between peaks is observable as well.

visualization confirms that these frequency channels between centers of highly overlapping peaks are not suitable for the normalization approach used in PGA.

5. Summary and outlook

This work shows the critical role of the peak-specific normalization within PGA for multivariate curve resolution. This approach is motivated using low-dimensional representations in U- and V-space and comparisons to AFS results. While standard PGA relies on an a priori peak selection, we introduced a new systematic frequency channel scanning approach that iterates through all available frequency channels to determine a normalization for the optimization of an objective function. Frequency channel scanning serves as a robust tool for identifying the most promising channels for spectral mixture analysis.

The autoPGA and frequency channel scanning algorithms are part of the Peak Group Analysis module in the FACPACK software package. The software is freely available at our website:

https://www.math.uni-rostock.de/facpack/

Possible future research topics related to the PGA include:

- The impact of different weighting factors on the frequency channel scanning results,
- the implementation of a scanning algorithm focused on the selection of weighting factors,
- a design of experiments consideration focused on the PGA, see [17],
- the implementation of simultaneous detection of multiple components in PGA
- the analysis of the impact that different constraints have on the convexity of the objective function,
- the implementation of comparisons to results from vibrational mode analysis based on quantum mechanical methods, such as Density Functional Theory (DFT), using the Wasserstein distance as described in [14] as part of the similarity constraint f_4 in the objective function.

Acknowledgement

Funding of Lukas Prestin and Esha Suresh Babu by the DFG Research Training Group 2943 "SPECTRE" (project number 507189291) is gratefully acknowledged.

References

[1] 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.

[2] J. Dennis, D. Gay, and R. Welsch. Algorithm 573: An adaptive nonlinear least-squares algorithm. *ACM Trans. Math. Softw.*, 7(3):369–383, 1981.

[3] 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.

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

[5] C. Kubis, M. König, B. N. Leidecker, D. Selent, H. Schröder, M. Sawall, W. Baumann, A. Spannenberg, A. Brächer, K. Neymeyr, R. Franke, and A. Börner. Interplay between catalyst complexes and dormant states: In situ spectroscopic investigations on a catalyst system for alkene hydroformylation. *ACS Catalysis*, 13(8):5245–5263, 04 2023.

[6] P. Lasch. Spectral pre-processing for biomedical vibrational spectroscopy and microspectroscopic imaging. *Chemom. Intell. Lab. Syst.*, 117:100–114, 2012.

[7] 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.

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

[9] E. R. Malinowski. Window factor analysis: Theoretical derivation and application to flow injection analysis data. *J. Chemom.*, 6(1):29–40, 1992.

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

[11] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.*, 11(2):431–441, 1963.

[12] K. Neymeyr, M. Beese, H. Abdollahi, and M. Sawall. Can angle measures be useful in MCR analyses? *J. Chemom.*, 38(11):e3582, 2024.

[13] K. Neymeyr, C. Kubis, L. Prestin, R. Franke, and M. Sawall. Why is the Peak Group Analysis so effective for IR spectra analysis? *J. Chemom.*, 40(4):e70121, 2026.

[14] K. Neymeyr, B. Zhang, C. Kubis, M. Brinker, L. Prestin, T. Andersons, M. Sawall, and J. Hellwig. A spectral matching algorithm based on the Wasserstein metric. *J. Chemom.*, 40(6):e70131, 2026.

[15] M. Sawall, C. Kubis, E. Barsch, D. Selent, A. Börner, and K. Neymeyr. Peak group analysis for the extraction of pure component spectra. *J. Iran. Chem. Soc.*, 13(2):191–205, 2016.

[16] 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.

[17] 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.*, 34(2):e3159, 2020.

[18] M. Sawall, C. Kubis, D. Selent, A. Börner, and K. Neymeyr. A fast polygon inflation algorithm to compute the area of feasible

solutions for three-component systems. I: Concepts and applications. *J. Chemom.*, 27(5):106–116, 2013.

[19] M. Sawall, C. Ruckebusch, M. Beese, R. Francke, A. Prudlik, and K. Neymeyr. An active constraint approach to identify essential spectral information in noisy data. *Anal. Chim. Acta*, 1233:340448, 2022.

[20] 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: Chemical and Biochemical Data Analysis*, pages 199–231. Elsevier, 2020.

[21] G. W. Stewart. Determining rank in the presence of error. Technical report, University of Maryland, College Park, MD, USA, 1992.

[22] K. X. Wan, I. Vidavsky, and M. L. Gross. Comparing similar spectra: from similarity index to spectral contrast angle. *J. Am. Soc. Mass Spectrom.*, 13(1):85–88, 2002.

[23] E. Widjaja, C. Li, W. Chew, and M. Garland. Band target entropy minimization. A robust algorithm for pure component spectral recovery. Application to complex randomized mixtures of six components. *Anal. Chem.*, 75(17):4499–4507, 2003.

[24] W. Windig and J. Guilment. Interactive self-modeling mixture analysis. *Anal. Chem.*, 63(14):1425–1432, 1991.

[25] H. Zhang, W. Chew, and M. Garland. The multi-reconstruction entropy minimization method: Unsupervised spectral reconstruction of pure components from mixture spectra, without the use of a priori information. *Appl. Spectrosc.*, 61(12):491–498, 2007.